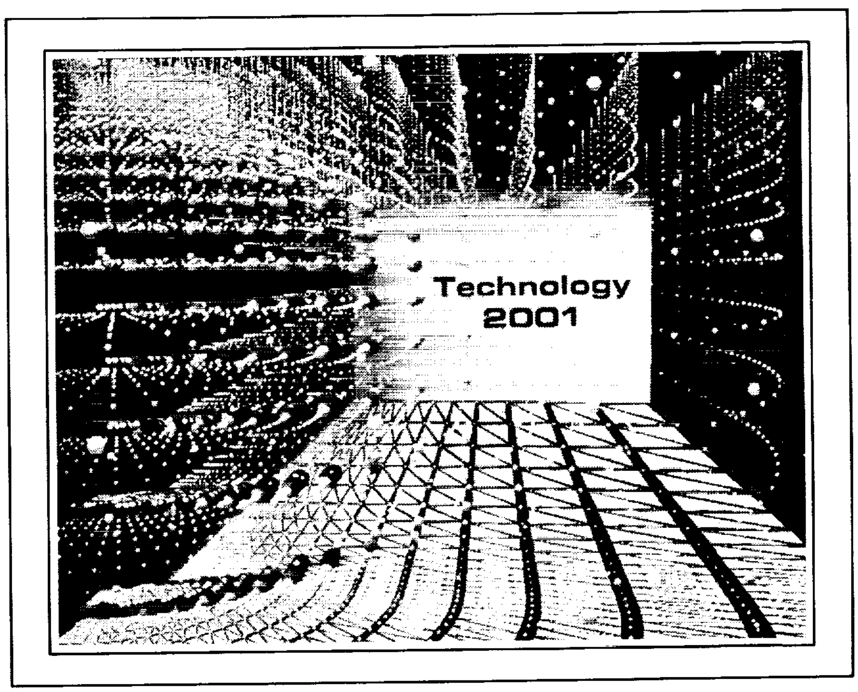
NASA Conference Publication 3136, Vol. 1

TECHNOLOGY 2001

The Second National Technology Transfer Conference and Exposition

December 3-5, 1991
San Jose Convention Center
San Jose, CA



Conference Proceedings

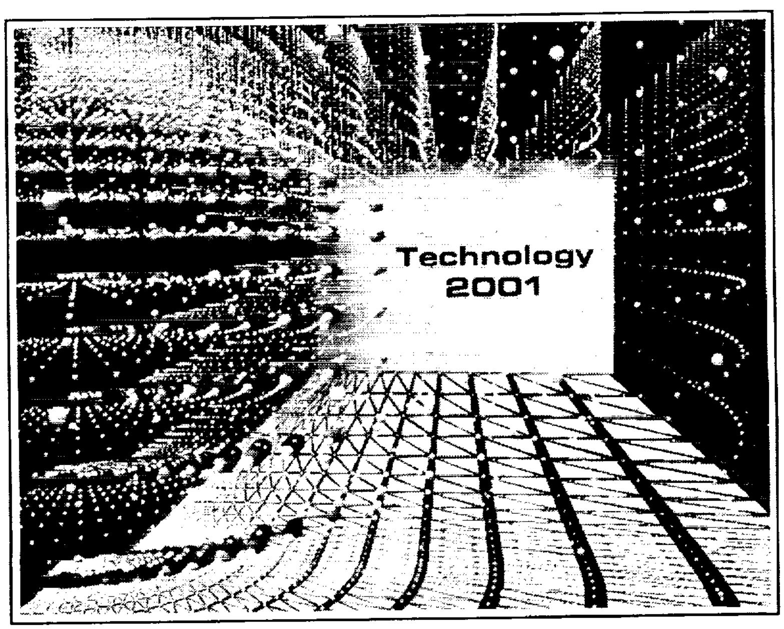
Sponsored by NASA, the Technology Utilization Foundation, and
NASA Tech Briefs Magazine

NASA Conference Publication 3136, Vol. 1

TECHNOLOGY 2001

The Second National Technology Transfer Conference and Exposition

December 3-5, 1991
San Jose Convention Center
San Jose, CA



Conference Proceedings

Sponsored by NASA, the Technology Utilization Foundation, and *NASA Tech Briefs Magazine*

TECHNOLOGY 2001 - SYMPOSIA PROCEEDINGS

Presented December 3-5, 1991
San Jose, California

TECHNOLOGY 2001 was the second national technology transfer conference and exposition. Held at the San Jose Convention Center December 3-5, 1991, TECHNOLOGY 2001 built upon the foundation laid by last year's initial conference in Washington, D.C., the mission being to transfer advanced technologies developed by the Federal government, its contractors, and other high-tech organizations to U.S. industries for their use in developing new or improved products and processes.

TECHNOLOGY 2001 was sponsored by the National Aeronautics and Space Administration (NASA), *NASA Tech Briefs* magazine, and the Technology Utilization Foundation, with the participation of the following Federal agencies:

Department of Agriculture	Department of Commerce
Department of Defense	Department of Energy
Department of Health and Human Services	Department of the Interior
Department of Transportation	Department of Veteran Affairs
Environmental Protection Agency	National Science Foundation

In addition to an exhibit showcasing the products and technologies available for sale or license from over 200 exhibitors, this year's conference featured 30 concurrent technical sessions in which 120 papers were presented, agency workshops, industry briefings, and the annual Intelligent Processing Equipment (IPE) Conference held concurrently with TECHNOLOGY 2001.

We are pleased to provide the proceedings from the 30 concurrent sessions. This year's program featured symposia on Advanced Manufacturing, Artificial Intelligence, Biotechnology, Computer Graphics and Simulation, Communications, Data and Information Management, Electronics, Electro-Optics, Environmental Technology, Life Sciences, Materials Science, Medical Advances, Robotics, Software Engineering, and Test and Measurement.

The proceedings have been published in two volumes. Volume One contains the first 60 papers presented (in order), while Volume Two contains the last 60 papers presented (again, in order). Proceedings from the IPE Conference Symposia are published under separate cover.

This is Volume One. Again, the papers appear in the order in which they were presented at TECHNOLOGY 2001. For information regarding additional copies, please contact:

THE TECHNOLOGY UTILIZATION FOUNDATION
41 East 42nd Street, #921
New York, NY 10017

TECHNOLOGY 2001 SYMPOSIA PROCEEDINGS - VOLUME ONE

Table Of Contents

Electronics:

Environmental Technology:

Materials Science:

Medical Advances - Computers in Medicine:

Software Engineering:

Data and Information Management:

Electro-Optics:

Life Sciences:

ADVANCED MANUFACTURING

(Session A1/Room A1)

Tuesday December 3, 1991

- Ceramic Susceptor for Induction Bonding of Metals, Ceramics, and Plastics
- Applying NASA's Explosive Seam Welding
- Laser-Based Weld Joint Tracking System
- Precision Joining Center

N92-22424

CERAMIC SUSCEPTOR FOR INDUCTION BONDING OF METALS, CERAMICS, AND PLASTICS

Robert L. Fox
Langley Research Center
Hampton, VA

John D. Buckley
Langley Research Center
Hampton, VA

ABSTRACT

A thin (.005) flexible ceramic susceptor (carbon) has been discovered. It was developed to join ceramics, plastics, metals, and combinations of these materials using a unique induction heating process. Bonding times for laboratory specimens comparing state of the art technology to induction bonding have been cut by a factor of 10 to 100 times. This novel type of carbon susceptor allows for applying heat directly and only to the bondline without heating the entire structure, supports, and fixtures of a bonding assembly. The ceramic (carbon film) susceptor produces molten adhesive or matrix material at the bond interface. This molten material flows through the perforated susceptor producing a fusion between the two parts to be joined, which in many instances has proven to be stronger than the parent material. Bonding can be accomplished in 2 minutes on areas submitted to the inductive heating. Because a carbon susceptor is used in bonding carbon fiber reinforced plastics and ceramics, there is no radar signature or return making it an ideal process for joining advanced aerospace composite structures.

INTRODUCTION

Induction heated Rapid Adhesive Bonding (RAB) techniques using a non-metallic susceptor for joining plastics, metals, and ceramics have been developed at the Langley Research Center (LaRC) (1). This process permits the heating of thermoset adhesive filled susceptors or the interface of thermoplastics directly at the bondline. Rapid Adhesive Bonding involves an electromagnetic induction heating of thin ceramic material (carbon susceptor) embedded in the bondline of the structure (Figure 1). Because only the bondline and material in the immediate area are heated, thermal distortions are less severe than conventional processes, which simplify and lower the cost of fixturing. Heating rates greater than 600_F in 30 seconds have been generated employing a recently discovered ceramic (graphite) susceptor using RAB procedures (Figure 2). RAB bonds have been produced in less than 2 minutes, consuming much less power than conventional techniques. The low amount of input electrical power required to heat the bondline can be supplied from various sources (Figure 1).

Current state-of-the-art processes, such as press or autoclave bonding, take hours to accomplish and have very limited heating/cooling rate capabilities. These current processes rely on the conduction of heat from resistance heating elements through tooling, fixtures, caul plates, the structural parts and finally into the bondline to heat the adhesive. Consequently, a heat-up rate of 10_F/min is considered high, and much energy is consumed in bonding structures together. These bonding cycles can often take over 5 hours to execute.

The original objective of this bonding system was to provide low energy, portable, selfcontained, cost-effective apparatus and method for joining thermoplastic matrix composites and other compatible materials. This equipment was developed to fabricate structures to be used in outer space, and secondarily, structures on earth, or in motionless surroundings. As stated above, a recently discovered ceramic (graphite) susceptor material has been used to join pieces of metallic, ceramic, and plastic composites. In a toroid pole piece, magnetic flux remains inside the toroid core when the system is energized. To divert the path of the magnetic flux from the toroid to an adjacent ceramic susceptor, the toroid must be altered. This alteration is accomplished by cutting a segment out of the toroid and placing the air gap in the toroid on the surface of a matrix material composite sandwich consisting of a susceptor positioned between the two composite components to be joined (Figure 1). When using inductive heating to bond a typical plastic composite, a toroid is first energized, flux will flow through the toroid, through the plastic composite (which is transparent to magnetic flux) into the ceramic susceptor back through the plastic composite into the toroid. Alternating current produces inductive heating instantly in the susceptor causing the plastic interfacing on either side of the susceptor to melt and flow into perforations made in the ceramic susceptor forming the joint. Joining is accomplished in minutes.

The objective of this proof of concept study was to demonstrate the thermal efficient quality of a ceramic (graphite) susceptor when used for the induction heating and subsequent joining or bonding of plastic composites, metals, ceramics, and combinations of these materials.

SPECIMEN PREPARATION

Components of the specimens are shown in Figure 3 laid out in the order in which they would be stacked together in the fixture. A susceptor is sandwiched between thermoplastic adherends or in a stack containing adhesive layers placed between a thermoset plastic or between inorganic adherends (metal or ceramic). The surface preparation for all lap shear specimens consisted of a methanol wash followed by a 120 grit sandblast plus a second wash in acetone, methanol, and trichloroethylene. Table 1 shows the materials used in this ceramic susceptor proof of concept study.

BONDING AND TESTING

Shear Specimen Bonding

Overlap shear specimens were bonded in a configuration conforming to the American Society for Testing Materials (ASTM) standards D1002 and D3136. The technique similar to that used for spot welding metallic structures was used for rapid bonding of lap shear specimens made of thermoplastic composites, thermoset composites, metals, ceramics, and combination of these materials (Figure 4).

The rapid bonding equipment for laboratory shear specimens is shown in Figures 5 & 6. The press is identical to that for conventional specimen bonding, as are the load cell and temperature and load indicators. Replacing the conventional heated platens is a toroidal high frequency induction heater and its power controller. The specimen is located in a fixture for ease of alignment. The fixture was fabricated to align the specimen components prior to bonding. It was machined from bakelite with cutouts and location screws for the adherend (Figure 6). Bonding was accomplished by assembling the specimen in the specimen fixture, placing the fixture in the press under the toroid head and applying pressure and the induction field. The power used to energize the induction heater was approximately 300 watts at 60 Hz and 120 volts input into the inductive heater circuit. When power was applied, the induced energy from the toroid rapidly heated a perforated graphite susceptor which had been impregnated with a thermoplastic adhesive or was sandwiched between thermosetting adhesive films. The power was concentrated as heat entirely within the ceramic (graphite) susceptor, concentrating the heat within the bond line and minimizing detrimental thermal effects on the composite shear test specimen. For lap-shear specimens, the ceramic, metallic, or fiber-reinforced plastic composite material adherends were placed above and below the susceptor in the specimen fixture, and bonding pressure was applied (Figure 1). The susceptor heated the adhesive or thermoplastic composite adherend rapidly, usually within a minute, to the bonding temperature. Temperature within the bondline was considered to be an important requirement of this induction bonding process since heating was concentrated in the bondline in all applications in which a susceptor was used. A thermocouple was positioned in the bondline of each test specimen for each of the materials to be bonded (Figure 6). The heat is maintained from one to several minutes to promote adherend joining. When power is turned off, the specimen rapidly cools to a temperature below which the adhesive or thermoplastic composite is sufficiently set, and pressure is removed. Some of the composite materials tested are shown in Table 1. This process is more controllable and more energy conserving than conventional bonding with heated platens or an autoclave. (1,2).

APPARATUS AND TEST PROCEDURES

Tensile tests at room temperature were conducted in a 10 kilo pounds mechanical power screw driven machine at a head speed of .05 inches per minute until fracture. Grips used in the tensile tests were split collar assemblies. Maximum load was recorded from the dial indicator on the test machine, recorded from the dial indicator on the test machine. Specimen shear area used to determine shear strength was accomplished by measuring and taking the sum of all the hole areas in the graphite susceptor sandwiched between the adherend tensile specimen (Figure 7).

DISCUSSION OF RESULTS

Graphite-Peek Adherends

Table 1 and Figure 8 show overlap shear strengths (per ASTM D1002) of graphite fiber polyether etherketone (PEEK) (.004 inches thick) fabricated by rapid adhesive bonding technique using PEEK adhesive and a 0.0005 inch perforated ceramic (carbon) susceptor (Figure 3). Data are shown for specimens bonded at 720^0 F at 32 psi. Hold time under pressure at the bond temperature was 2 minutes. All shear strength data was obtained at room temperature. Figure 8 shows the best shear strength value obtained joining graphite/PEEK to graphite/PEEK with PEEK WAS 4,500 PSI. The bond was cohesive through the perforated carbon susceptors and failure was observed in the adherend part of the tensile test specimen.

Titanium Adherends

Table I and Figure 9 show the overlap shear strengths (per ASTM D1002) of Ti-6AL-4V titanium alloy adherends (0.05 inch thick) fabricated employing RAB using PEEK thermoplastic adhesive and a 0.0005 inch thick perforated ceramic (carbon) susceptor sandwiched with adhesive (similar to Figure 3). The specimens were bonded at a temperature of 720^0 F and 32 psi. Hold time under pressure at the bond temperature was 2 minutes. All shear strength data was obtained at room temperature. The highest shear strength value obtained joining the titanium adherends was 6,500 psi (Figure 9). The bond material was PEEK adhesive joining the two adherend components through a perforated ceramic (carbon) susceptor at the joint interface. The failure of this specimen was in the adhesive bond.

Titanium was also bonded to titanium using Hysol EC934 thermoset adhesive. The titanium adherends were 0.05 inches thick with a .005 inch thick ceramic (carbon) susceptor filled with the Hysol EC934 adhesive and sandwiched between the two adherends that made up the shear specimen. Table 1 and Figure 10 show the shear strength for this combination of material. The specimens were bonded at a temperature of 400^0 F and a bonding pressure of 19.2 psi. Hold time under pressure at the bond temperature was 2 minutes. All shear strength data was obtained at room temperature. The highest adhesive bond strength using Hysol EC934 adhesive was 6,400 psi. The failure of this specimen was in the adhesive.

Graphite/Epoxy Adherends

The applicability of a ceramic (carbon) susceptor used with RAB to bond graphite/epoxy adherends was demonstrated when joining Hercules 350 graphite/epoxy thermoset adherend with Hysol EC934 thermoset adhesive. Specimens were bonded at a temperature of 400^0 F and a pressure of 19. Hold time under pressure at the designated temperature of 2 minutes (Table 1, Figure 11). The highest shear strength value obtained bonding thermoset composite to itself with Hysol EC934 thermoset adhesive was 2,250 psi (Figure 11). It was noted that the bond maintained its integrity, and failure occurred in the adherend.

Adherends of Unlike Materials

The versatility of the rapid bonding concept was again demonstrated by using the process described in the preceeding paragraphs to bond titanium to graphite epoxy, aluminum to graphite epoxy, and aluminum to aluminum oxide ceramic. The adhesive used was Hysol EC934. The bonding temperature was 400^0 F. The average tensile strength of the 4,500 psi as shown in Table 1 and Figures 12, 13, and 14. Figure 12 shows strength data for Titanium bonded to graphite epoxy composite with Hysol EC934. The best shear strength value obtained for these specimens was 2,900 psi. Failure of the specimen was in the composite material. Aluminum 6061-T6 was bonded to graphite epoxy Hercules 3501 with Hysol EC934. The fabrication parameters and procedures for bonding the aluminum to the graphite epoxy was the same as described earlier in the text (Table 1). The average strength of the specimens tested was 432 > psi and the best strength value for the combination of materials was about 5250 psi. The 3501 adherend epoxy is .061 thick and the aluminum is .062 inches thick. The specimens failed in the composite part of the overlap shear joint. The last combination of materials bonded together was aluminum 6061-T6 (.062 inches thick) and aluminum oxide (.062 inches thick). The perforated ceramic (carbon) susceptor (.005 inches thick) was filled with Hysol EC934 thermoset adhesive and sandwiched at the joint between the aluminum and aluminum oxide adherends. Upon completion of the bonding cycle the specimen was tensile tested and found to have an average strength of 4,520 psi. The best shear strength value obtained from the aluminum oxide test was 5,600 psi. Failure occurred in the ceramic portion of the specimen. The low numbers obtained when testing this group of specimens is believed to be due to the lack of mobility in the grips of the pull test machine and the brittle nature of the aluminum oxide ceramic.

CONCLUDING REMARKS

A proof of concept study at the Langley Research Center has been conducted to evaluate a ceramic (carbon) susceptor for use in the induction bonding of structural materials used in aerospace technology. A thin (.005) flexible ceramic susceptor (carbon) has been developed to be used with a toroid bonder inductive heating instrument. Preliminary tests show that this bonding process produces rapid joining of ceramics, plastics, metals, and combinations of these materials. A typical lap-shear specimen placed in the toroid inductive heating press produced a bond in less than 10 minutes from energizing to removal from the heating press. Average lap shear bond strengths varied from about 6,000 psi to 2,000 psi depending on the materials bonded. Some specimens failed in the adherend rather than the bond joint. Bonding times for laboratory specimens comparing state-of-the-art technology to induction bonding have been cut by a factor of 10 to 100 times.

REFERENCES

1. Stein, B. A.; Tyeryar, J. R.; and Hodges, W. I.: Rapid Adhesive Bonding Concepts. NASA TM 86256, June 1980.

2. Buckley, J. D.; Fox, R. L.; and Swaim, R. J.: Toroid Joining Gun SAE Paper 85040T presented at International Congress and Exposition Detroit , MI, February 25- March 1, 1985.

TABLE I

SPECIMEN BONDING PARAMETERS

ADHEREND	[1]GR PEEK TO GR/PEEK	TI to 3501[2]	AL TO 3501[2]	TI TO TI	AL TO AL2O3	TI TO TI	3501[2] 3501[2]
ADHESIVE	PEEK	HYSOL934	HYSOL934	HYSOL934	HYSOL934	HYSOL934	HYSOL934
SUSCEPTOR	CARBON	CARBON	CARBON	CARBON	CARBON	CARBON	CARBON
BONDING TEMP DEGREES F	720	400	400	720	400	400	400
BONDING PRES PSI	32	19.2	19.2	32	19.2	19.2	19.2
SURFACE PREP	A	A	A	A	A	A	A
HOLDING TIME AT TEMP MIN	2	2	2	2	2	2	2
AVG OVERLAP SHEAR STRENGTH PSI	4371	2566	4327	5993	4520	5646	2022

A. METHANOL WASH, 120 GRIT BLAST, ACETONE, METHANOL AND
1. GRAPHITE - POLYETHERTHER-KEPTON COMPOSITE, TRICHLOROETHLYENE
2. HERCULES THERMOSET GRAPHITE EPOXY

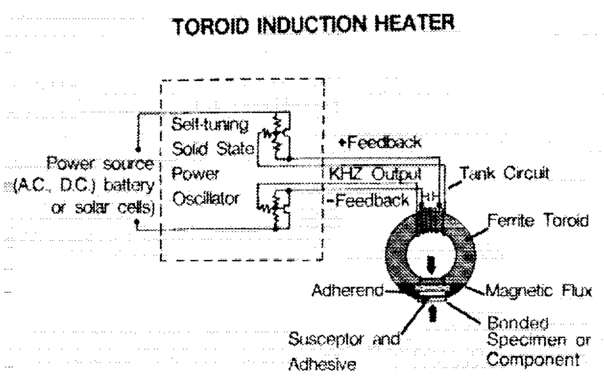


Figure 1 Schematic of toroid induction concept

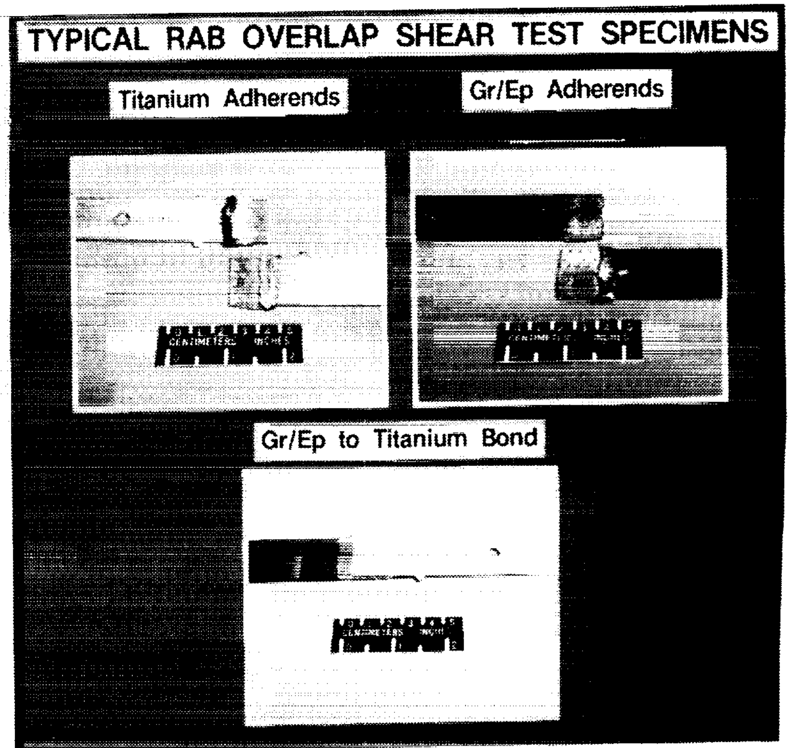


Figure 4 Overlap shear specimans according to ASTM

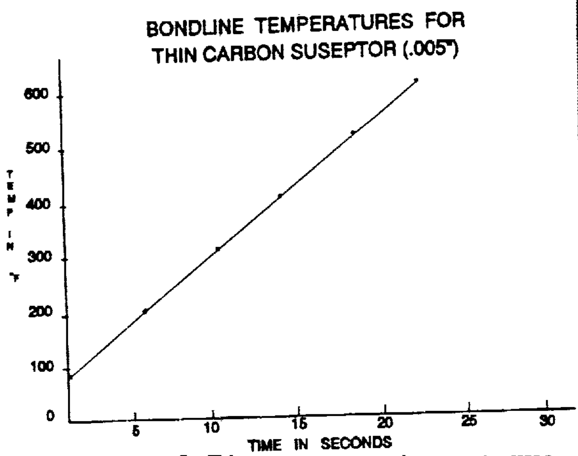


Figure 2 Time temperature curve for inductive heating of a ceramic (graphite) susceptor

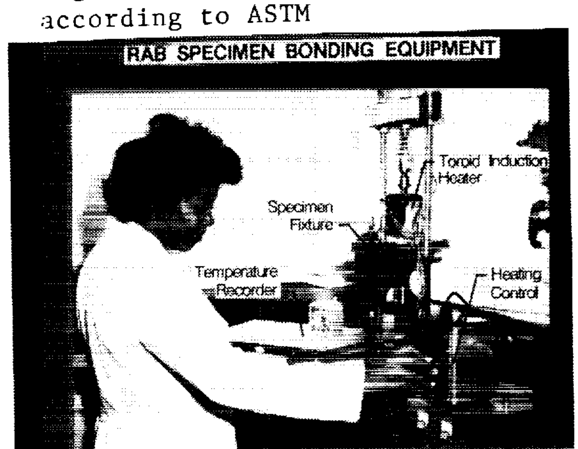


Figure 5 Toroid induction heater speciman bonding equip●ment

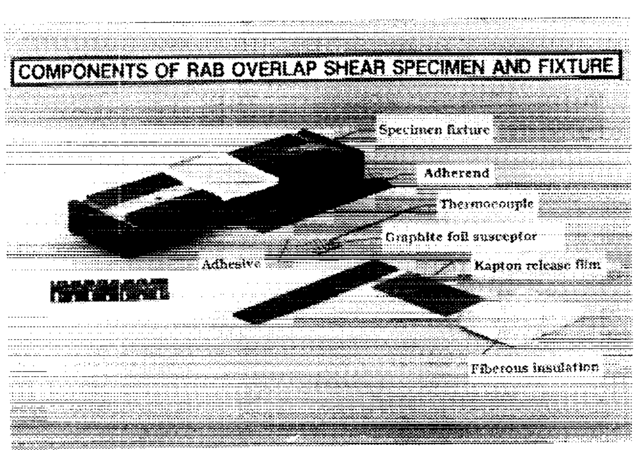


Figure 3 Typical induction bonded overlap shear test specimams

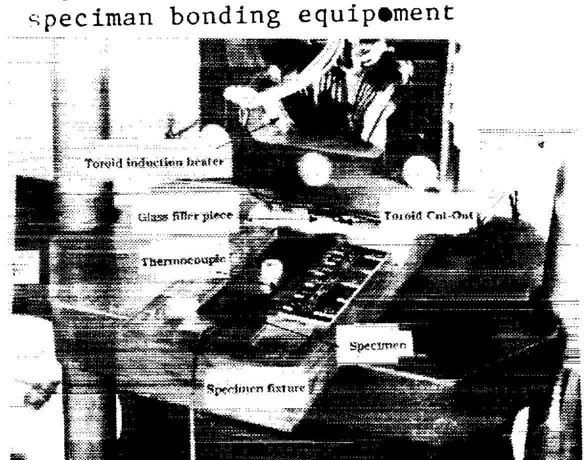


Figure 6 Speciman fixture in press under toroid head

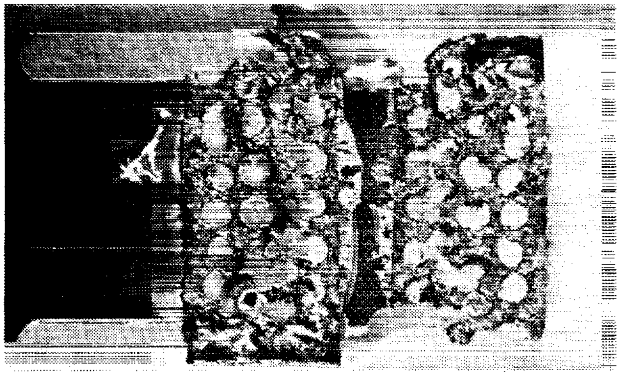


Figure 7 Overlap shear test speciman showing adhesive through preforated susceptor(both sides).

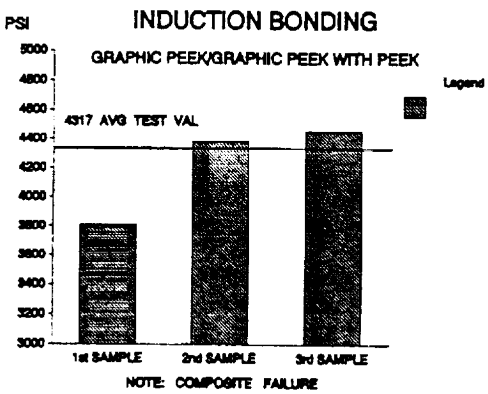


Figure 8

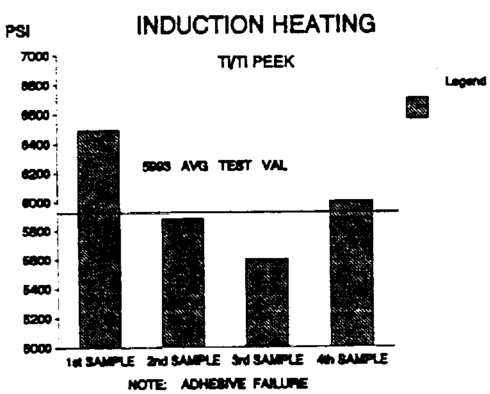


Figure 9

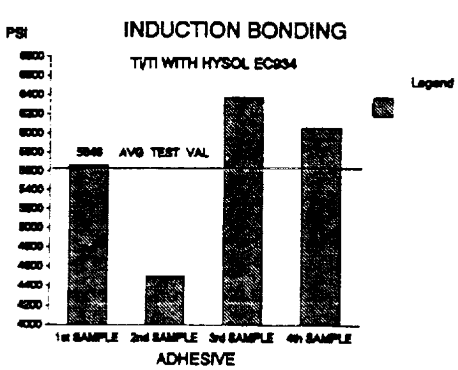


Figure 10

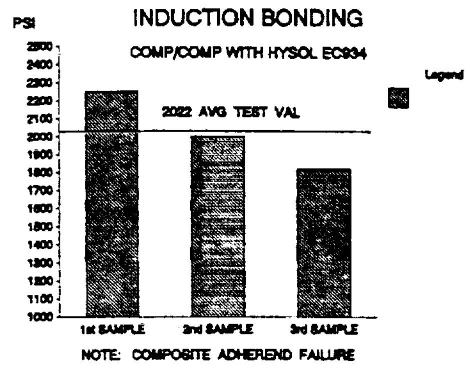


Figure 11

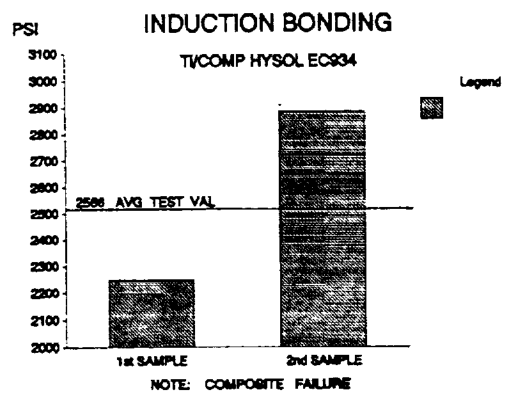


Figure 12

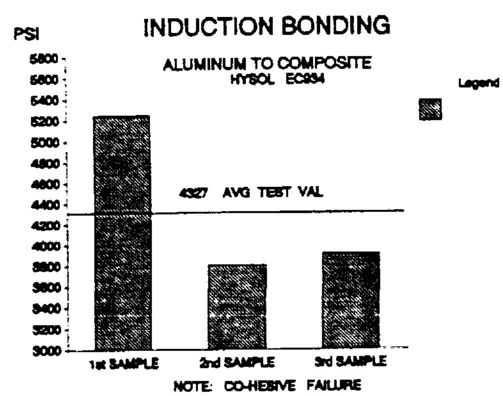


Figure 13

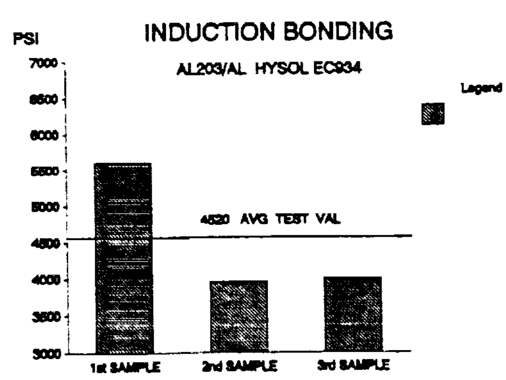


Figure 14

N92-22425

APPLYING NASA'S EXPLOSIVE SEAM WELDING

Laurence J. Bement
NASA Langley Research Center
Hampton, Virginia

ABSTRACT

This paper summarizes the status of a novel explosive seam welding process, invented at NASA's Langley Research Center in the 1960's, developed and evaluated for a wide range of metal joining opportunities and now being used commercially. The process employs very small quantities of explosive in a ribbon configuration to accelerate a long-length, narrow area of sheet stock into a high-velocity, angular impact against a second sheet. At impact, the oxide films of both surfaces are broken up and ejected by the closing angle to allow atoms to bond through the sharing of valence electrons. This cold-working process produces joints having parent metal properties, allowing a variety of joints to be fabricated that achieve full strength of the metals employed. Successful joining has been accomplished in all aluminum alloys, a wide variety of iron and steel alloys, copper, brass, titanium, tantalum, zirconium, niobium, telerium and columbium. Safety issues have been addressed and are as manageable as many currently accepted joining processes. The Atomic Energy of Canada is evaluating this process for zirconium attachments in nuclear reactors. DEMEX International has licensed a NASA tube-joining patent and is applying it to tube plugging.

INTRODUCTION/BACKGROUND

Although the demand is increasing for highly reliable, metal joining processes both for hazardous or inaccessible operations and for the applications of advanced metal alloys and combinations, there is a general reluctance to accept explosive welding processes because of a perception that explosive materials cannot be safely managed. This combination of safety concern and the resistance to accepting novel joining methods has virtually reduced the potential user community to those that have exhausted all other alternatives. The purpose of this paper is to compile the history of the NASA explosive seam welding technology and provide the logic to show the practicality of its application in modern joining requirements. This will be accomplished by introducing the principles of explosive joining and the attendant variables, and presenting the NASA explosive seam welding process in terms of types of joints created, capabilities, the safety issue, and the applications considered for its use.

EXPLOSIVE JOINING PRINCIPLES

Explosive welding is a cold-working process that produces metallurgical bonds, exhibiting parent metal properties, which are impossible to achieve by any other joining process. The explosive welding process is accomplished by a high-velocity, angular collision of metal plates, which effaces the oxide films on both surfaces to allow interatomic (electron sharing) linkups. See references 1 through 4. The angular collision and parameters are shown in the top sketch of figure 1. The several thousand megapascal (several million psi) explosive pressure drives the flyer plate (in a near-fluid condition) to velocities of a thousand m/sec (3.28 thousand ft/sec). On impact, the kinetic energy is converted to skin-deep (less than 0.0025 cm (0.001 in)) melts, which are stripped from the surfaces and squeezed out by the closing angle. Two explosive joining processes now exist, cladding and seam welding.

The explosive cladding process (reference 1) utilizes bulk explosive, such as dynamite or nitroguanidine, to create an explosive pressure input that travels at a velocity of approximately 1,300 to 3,300 m/sec (4,000 to 10,000 ft/sec) to create the angular collision. For the lead-to-steel cladding process in reference 1, 79 kg (175 lbs) of loose-powder dynamite is literally shoveled onto the 1.2 x 4.9-m (4 x 8-ft), 0.3-cm (0.125-in) thick flyer plate, which is spaced in parallel to the base plate. The explosive is detonated along the 4.9-m (8-ft) edge. Explosive cladding is limited to lengths of approximately 4.9 m (10 ft), due to the inability to maintain the collision parameters. The resulting bonds exhibit parent metal properties, and are generally observed through polishing and etching as the wavy weld interface, indicated in the top sketch in figure 1. These waves are truly, "skin deep," usually less than 0.006 cm (0.002 in), peak-to-peak; a stainless steel flyer plate, 0.0025 cm (0.001 in) in thickness, has been successfully joined and exhibited the full strength of the parent foil. Each set of variables, described later, produces a "signature" interface, which is unique to that set of variables. This signature can be virtually free of waves to very deep

(0.013 cm or 0.005 in), complex patterns. Microscopic analyses of this interface has indicated some entrapment of oxides and a slight amount of work hardening, but the actual interface of the two plates is an indefinable line, only as pronounced as the grain boundaries in the metals themselves. Figure 2 shows an example of an explosively joined interface, that is impossible to achieve by any other joining process, aluminum alloys of 2024-T4 to 6061-T6.

The NASA explosive seam welding process (references 2 through 9) differs from cladding in the explosive used and the angular collision mechanisms. The explosive used is cyclotrimethylene-trinitramine (RDX), which is encased in a lead-sheathed "ribbon," as shown in table I. The explosive load is measured in gr/ft (0.0198 g/m) with 7,000 gr/lb, and has a velocity of explosive propagation of 7,900 m/sec (26,000 ft/sec). The plates are initially separated by approximately 0.04 cm (0.015 in), and the ribbon explosive is taped to the flyer plate. On initiation of the explosive, the center portion is driven downward, as shown in the lower sketches in figure 1, to produce the high-velocity angular impact, from the center outward to both sides at a 60-degree "jet" angle. The resulting joint is highly uniform and just under the width of the ribbon explosive selected to accomplish the joining. As a comparison of efficiency to the cladding process described above, a 79-kg (175-lb) quantity of RDX could produce a continuous joint in 0.32-cm (0.125-in) thick aluminum 15,000 m (49,000 ft) in length. With an approximate bond width of 0.64 cm (0.25 in), the total bond area would be over 94 m^2 (1020 ft^2), as compared to the 3 m^2 (32 ft^2) in the cladding process.

EXPLOSIVE SEAM WELDING VARIABLES

The following variables must be optimized for every joining configuration, as described in references 2 through 4.

1. Plate material
2. Plate thickness
3. Explosive quantity
4. Standoff (plate separation)
5. Surface finish and cleanliness
6. Mechanical shock

Metal alloy, condition and thickness present a wide range of density, mass, hardness, strength, rigidity and malleability. These variables directly influence the quantity of explosive necessary to bend and accelerate the plates to achieve explosive joining. As metal density, mass, strength and rigidity increase, the explosive quantity must be increased in a non-linear progression.

Standoff, or separation between the plates, is also required for the joining mechanism; standoffs of 0.025 to 0.064 cm (0.010 to 0.025 in) are readily achievable by means such as shims, tape, or fixturing. Larger standoffs not only reduce the efficiency of the joining operation, but also introduce fracturing of internal grain boundaries. A notch can be machined in the surface of either or both plates to achieve the necessary separation. The plates can be configured to present a parallel or angular interface. Prebending one or both plates efficiently introduces the necessary collision angle, which results in a larger bond area. The same result is achieved by machining a V-shaped notch, as shown in a later example.

Surface cleanliness and smoothness must be carefully managed to achieve explosive joining success. The properties of substantial amounts of oxide films, such as rust or the thinner, harder and tougher oxide on aluminum, as well as water, grease or oil, prevent explosive joining. Low-carbon iron alloys must be degreased and polished to remove corrosion-protective greases and mill scale (etching is not recommended, since rusting becomes extremely rapid). Stainless steel alloys need only degreasing and a final alcohol wipe. Pure aluminum has a minimal oxide film, requiring only degreasing. However, the oxide films on aluminum alloys require chemical etching for removal to allow reliable joining under reasonable, noncorrosive ambient conditions over a several-week time frame. Since explosive joining is a "skin-deep" process, surface finishes more than 0.008 cm (0.003 inch) in depth prevent joining; a surface finish of 32 rms, which is rougher than virtually all sheet metal stock, is adequate.

The mechanical shock generated by the explosive pressure used to accelerate the plates along with the shock generated by the impact of the plates are the most damaging influences in the explosive joining process. The relative amplitude and influence are dependent on materials and structural configuration. These shock waves can not only damage sensitive structure in the area of the process, but can actually destroy a bonded joint immediately after its creation. Shock waves can be reduced by placing additional structure in the area. This additional structure can be a plate on the opposite side of a joining process (anvil), or clamping anvils in the immediate area. Adequate shock absorption can be achieved in the structure to be joined, particularly with thicker base plate materials.

TYPES OF JOINTS

Four different types of lap joints have been demonstrated, as shown in figure 3. The dissimilar-thickness joining process was described in the bottom illustration of figure 1. Similar-thickness joints can be achieved with the explosive ribbon on one side, but a more reliable approach places explosive ribbons on both sides of the separated plates. The ribbons are simultaneously initiated from a single source, such as a blasting cap, thereby balancing the explosive pressure waves. The sandwiched butt joint combines the above two approaches. The scarf joint (reference 5) is created by shifting the longitudinal axes of the explosive ribbon to create unbalanced forces. The plates are bent into axial alignment and joined in a single operation. These basic joints can be applied in a variety of configurations, such as curved surfaces and tubes.

CAPABILITIES

The NASA explosive seam welding process has many capabilities which are comparable or superior to currently accepted joining processes. The following are summaries from the detailed descriptions found in reference 4.

<u>Simple, high performance</u> - This process is simple, requiring little material preparation and tooling, while producing high-strength, hermetically sealed, fatigue-resistant joints. Once the explosive joining parameters have been established, the setup becomes purely mechanical with minimal requirements for personnel training and certification. Explosive joining requires comparable preparations to currently accepted joining processes. Complex tooling to assure heat sinking and the prevention of material distortion as required for fusion welding are unnecessary; the explosive ribbon need only to be taped in place. Joints that exhibit full strength of the stock metals and parent properties throughout the bond will be achieved with explosive seam welding by proper selection of joining parameters. Since the joining process bonds at the atomic level, the joints achieve hermetic seals (described later in this text). The fatigue resistance of a 6061-T6 explosively welded lap joint (reference 6) was superior to a high-efficiency fusion welded butt joint, in spite of the asymmetry of the lap joint; the explosively welded joint was in fact comparable to the parent stock.

<u>Thin to thick joints</u> - This process can join very thin materials to very thick materials. Better results are achieved by joining thin to very thick materials, due to dissipation of mechanical shock.

<u>Variety of alloys and combinations</u> - A wide variety of metals, alloys and combinations have been demonstrated with this process. Table II lists the metals and range of thicknesses in which 100 percent strength joints have been achieved. Table III lists the combinations of metals that have been joined, again achieving full strength of the weaker of the two alloys.

<u>Inspectable</u> - These joints can be thoroughly inspected to assure complete integrity by using nondestructive ultrasonic methods. Since the surfaces and thicknesses of the resulting joints are highly uniform, the bond areas can be precisely located and evaluated.

<u>No long-length limitations</u> - Although an approximate 0.64-cm (0.25-inch) initial length is required to stabilize this welding process, there are no long-length limitations for explosive seam welding. The explosive ribbon can be manufactured in lengths of 100 m (328 ft), and can be spliced.

<u>Remote joining capability</u> - Explosive seam welding is ideally suited for remote operations and potentially hazardous conditions. This process has the potential for hands-on to deep-space operations. A totally confined process is described in reference 7. Placing the ribbon explosive inside a flattened steel tube, fully contains all explosive products to prevent harm to personnel or surrounding equipment. Once assembled and transferred to the use site, explosive initiation can be commanded by transmitters. This remote joining capability would be valuable for use in environments, such as nuclear radiation, toxic gases and extreme temperatures, that are hazardous to humans.

SAFETY

All safety aspects of the NASA explosive seam welding process can be controlled to levels comparable to currently accepted joining methods. Safety issues include the handling of the explosive materials, initiation and management of the explosive detonation products.

Explosive selection - Explosive materials are available that are insensitive to bullet impact and lightning, with demonstrated temperature stabilities to over 200 C (400 F) for 50 hours. These explosives used in the ribbon cannot be initiated by normal handling, such as cutting with scissors or a razor blade.

Initiation systems - Initiation systems must address both the prevention of inadvertent actuation, as well as reliable firing. Blasting caps, widely used industrial explosive initiators, are sensitive to a wide variety of extraneous inputs, such as impact, electrostatic voltages and stray electromagnetically induced or radio frequency energies. These potential hazards can be eliminated in a number of different ways; one approach is through the use of available exploding bridgewire (EBW) detonators rather than hot bridgewire blasting caps. Instead of a low-voltage/current input used for hot bridgewires, the EBW requires a unique high-voltage/current to "explode" a highly conductive bridgewire against insensitive explosive materials. Of course, interlocks and safing and arming systems for the actual firing system are still necessary. Mechanically initiated explosive transfer lines, used in the mining industry, are another approach.

Explosive containment - The greatest concern expressed by potential users are the products of the explosion; the energy in the pressure wave, the sound produced, the fragments and debris, and the smoke generated. The major advantage of the NASA explosive seam welding process is the use of very small amounts of explosive materials. For example, the joining of 0.32-cm (0.125-inch) aluminum requires 25 gr/ft (0.5 g/m) of ribbon explosive. A 6-m (20-ft) joint would require a total of 500 gr, or just over 28 g (1 oz) of explosive material. This small amount of explosive material actually generates few gas molecules; consequently, the actual pressure wave created decreases dramatically with distance from its original source (at a rate greater than the inverse distance, cubed). For example, within the first 30 cm (1 ft distance) away from the source, the pressure is less than 70,000 Pa (10 psi), and is much less than 7,000 Pa (1 psi) in the next 30 cm (1 ft). The debris created in the explosive joining process is primarily small-particle lead splatter and the tape used to position the explosive. The lead is easily captured by lightweight barriers, such as plywood. The residual airborne particles are primarily unreacted carbon, which can be collected in the same manner as arc welding fumes. Another approach for containment is to place the ribbon explosive inside a flattened steel tube, as described in reference 7; the tube fully contains all explosive products. In summary, properly selected explosive materials, initiators, firing systems and a 1-m (3.28 ft) width, height containment volume surrounding the length of explosive, or total confinement of the explosive at the source, will assure safe application of this joining process.

APPLICATIONAL EFFORTS

Several developmental efforts were made to apply this process in response to requests for support. One application is under evaluation, while another is now marketed commercially.

Sealing of Vessels

Three requests were received: the first for a method to close and seal a spacecraft vessel after collecting a sample from the surface of another planet, the second for a thin foil to seal an X-ray source and the third for a method to repair a puncture in the aluminum external tank for the Space Shuttle. All applications required joining a flat sheet to a plate.

Spacecraft vessel - For the spacecraft vessel, a 0.08-cm (0.032-in) thick, 6061-T6 aluminum disc was used, placing the explosive ribbon opposite the V-notched machined interface in the like-alloy base plate, shown in figure 4. Once joined, the vessel was helium leak-checked through a threaded port in the base plate with no leakage, pressurized to 0.7 MPa (100 psi) dry nitrogen, and again leak-checked with no leaks. A second 0.7 MPa (100 psi) pressurization caused the disc to burst, leaving the bond line completely intact.

X-ray source - The second request was to join a 0.0025-cm (0.001-in) thick, 300 series stainless steel foil to a 0.127-cm (0.050-in) thick, steel plate. A 5-cm (2-in) diameter bond line was accomplished in the foil, which could not be torn from the plate. This joint required the adhesive bonding of the foil to a 0.05-cm (0.020-in) thick sheet stock of aluminum, which increased the mass of the flyer plate and reduced the dynamics of the operation and prevented the crush-cutting of the foil. The adhesive shattered in the joining operation, debonding the foil from the aluminum and leaving the foil completely exposed.

Space Shuttle tank repair - The approach proposed for the third request, to repair a puncture in an aluminum vessel, is shown in figure 5. In this case, the two weld plates (patches) were prenotched to a 12.7-cm (5-in) diameter circle. Plates were to be placed on both sides of the skin to produce symmetrical loading during the joining operation. Two mild detonating fuses (MDF) transmitted the explosive initiation signal through simple flexible electrical conduit to the ribbon explosives. The explosive products would be contained by the firing box and acoustic chambers. The actual process was demonstrated with Space Shuttle 2219-T87 aluminum specimens. Again, full strength, hermetic seals were achieved.

Wire Splicing

The problem of achieving high-strength, fully conductive joining of wire to itself or to terminals is a universal problem. This explosive seam welding operation produces such joints, in the approach shown in figure 6. The solid wires are stripped of insulation and laquers, spread into a flat plane, alternating the wires from each side of the splice. A prebent copper (or other compatible metal) strip on which is mounted the ribbon explosive is slid over the wires and the explosive is initiated. There is no limit on the number of wires. Splicing of 0.063 and 0.23-cm (0.025 and 0.090-in) diameter copper wire has been demonstrated, using 0.076-cm (0.030-in) copper sheet stock. Once joined, the strip could be rolled and swaged into a smaller diameter. Since this is an atomic-level bond, conductivity would be expected to be very high.

Joining of Tubes and Strips to Tubes for Nuclear Reactors

Two collaborative efforts have been conducted with the Atomic Energy of Canada (AEC), a civil service organization responsible for technology development for the design and maintenance of the Candu nuclear reactor. The reactor generates high temperature/pressure water in 360, 10-cm (4-in) diameter, zirconium pressure tubes that are positioned horizontally through the reactor's containment (calandria) vessel and contain the fuel rods around which the water flows. These pressure tubes are contained within thin-walled, low-pressure (calandria) tubes that are sealed at the interior cylindrical faces of the calandria. Secondary 400 series stainless tubes are mounted on the outboard sides of the cylindrical calandria faces and adapted with fittings to interface to the pressure tube.

Large-diameter tube joint - The first request in 1981 was to develop a method to join a 20-cm (8-in) diameter, low-carbon steel sleeve on a bellows assembly to a similar steel adaptor flange, as shown in figure 7. The adaptor flange was machined to a 0.076-cm (0.030-in) thickness, V-notch interface. Full parent strength of the flange material was achieved, even when the adaptor flange was deliberately undersized by 0.15 cm (0.060 in) on the diameter. Direct joining of the low-carbon steel adaptor flange to the 403 stainless steel tube to replace a shrink-fit joint was also demonstrated. The implications of this effort are presented in reference 8. The use of this process could reduce the down time of the reactor and the radiation exposure to personnel 100 fold, compared to the currently accepted fusion welding process.

Strips to tube, exterior - The second effort, initiated in 1989, focused on joining 0.046-cm (0.018-in) thick zirconium strips to the outside of the 0.42-cm (0.165-in) wall thickness pressure tube to act as "ion getters" for the prevention of hydrogen embrittlement. The tubes were slipped over a close fitting mandrel to absorb the mechanical shock introduced into the tube, as well as to eliminate deformation of the tube. Considerable effort was made to protect both the tube and strips from damage from the explosive products. The results of these efforts are shown in figure 8. The AEC evaluation has shown excellent bonds, better than any other process they have examined. However, their metallurgists are concerned about stress lines that penetrate approximately 0.013 cm (0.005 in) into the tube. Long-term evaluations are now being conducted in which the joints are subjected to hydrogen-bearing compounds to accelerate the potential for hydrogen embrittlement.

Strip to tube, interior - The third effort applied the 0.046-cm (0.018-in) zirconium strips to the interior of the 0.127-cm (0.050-in) zirconium calandria tube to allow retention of ceramic spacers between the calandria and pressure tubes. An external mandrel was employed with dunnage and tape on the interior of the tube to capture the explosive debris. The requirements for this joining process are not as severe, since these tubes are not stressed as much as the pressure tubes. Samples are currently being evaluated by the AEC.

Small-Diameter Tube Joining

Two applications were evaluated, one to join tubes into fittings and the other to plug tubes.

Shuttle engine fitting - The liquid oxygen heat exchanger, located above the combustion chamber of the Space Shuttle's main engine, converts oxygen from a liquid to a gaseous state to pressurize the Shuttle's external tank. The assembly challenge was the joining of 316L steel tubing to Inconel 625, Incoloy 903 or Haynes 188 candidate materials for the end fittings. Requirements were temperatures from -162 to +427 C (-260 to +800 F) with the failure of the tubing being catastrophic. The effort was approached by using 30 gr/ft (0.6 g/m) of ribbon explosive to bond a 0.089-cm (0.035-in) thick 316L sheet stock to sample plates of the candidate materials, as shown in figure 9. No material indicated an advantage, since parent strength of the 316L was achieved in all three specimens. Figure 10 shows the assembly of the teflon tool (reference 9) used to join tubes to fittings. The small-diameter tube was loaded with an initiation charge, which was in turn initiated by the lead-sheathed explosive cord, projecting out of the centerline of the tool. The ribbon explosive was wrapped around the tool with its end butted into the initiator. Teflon tape was wrapped around the tool to assure a close fit with the tube. Figure 11 shows the joint achieved against a V-notched internal interface in a Haynes 188 fitting with a 0.66-cm (0.260-in) OD, 0.066-cm (0.026-in) wall tube. The upper figure shows the unsuccessful attempts to chisel/peel the tube out of the fitting.

Tube plugging - The above technology was adapted to tube plugging under a patent license to DEMEX International, as described in reference 10. Figure 12 shows the tools with external V-notches for several diameter tubes. The peak-to-peak ripple of the metallurgical bond, shown at lower right, is less than 0.005 cm (0.002 in). As stated in reference 10, in 1989 DEMEX and Southwestern Engineering Service Company "have made some 35,000 plug installations without an operational failure." The explosive welding technology , according to Southwestern Engineering, allows faster plugging, hence reduced down time, cuts plugging costs and increases reliability."

SUMMARY AND RECOMMENDATIONS

The NASA explosive seam welding process has demonstrated unique joining capabilities in sheet and tubular metal configurations, and has accumulated a history of consideration and acceptance that should provide potential users another fabrication option.

The small amounts of explosive used in this process produces narrow, controlled, full-strength lap joints in a variety of metal alloys and combinations. These joints exhibit hermetic seals, as well as resistance to fatigue.

This process not only can be managed safely in close proximity to personnel, but provides a capability for remote operation to eliminate the exposure of personnel to hazardous environments, such as from radioactive materials.

Successful demonstrations of this process in a variety of applications throughout the 20 years since its invention have shown its credibility. Considered were the sealing of vessels, wire splicing, joining of tubes and strips to tubes, joining of tubes to fittings and tube plugging. Its potential is being evaluated by the Atomic Energy of Canada for nuclear reactors. DEMEX, Incorporated has obtained a patent license from NASA for tube plugging.

With so many challenging joining problems in modern industry, as many joining methods as possible should be available; the NASA explosive seam welding process has achieved sufficient maturity to warrant its further consideration.

REFERENCES

1. Otto, H. E.; and Carpenter, S. H.: Explosive Cladding of Large Steel Plates with Lead. Welding Journal, pp 467-473, July 1972.

2. Bement, Laurence J.: Small-Scale Explosive Seam Welding. Presented at the Symposium on Welding, Bonding and Fastening, Williamsburg, Virginia, May 30 to June 1, 1972.

3. Bement, Laurence J.: Small-Scale Explosion Seam Welding. Welding Journal, pp. 147-154, March 1973.

4. Bement, Laurence J.: Practical Small-Scale Explosive Seam Welding. NASA TM 84649, 1983.

5. Bement, Laurence J.: Explosively Welded Scarf Joint. Patent 3,842,485, November 22, 1974.

6. Otto, H. E.; and Wittman, R.: Evaluation of NASA-Langley Research Center Explosion Seam Welding. NASA CR-2874, August 1977.

7. Bement, Laurence J.: Totally Confined Explosive Welding. Patent 3,797,098, March 19, 1974.

8. Aikens, A. E.; and Bement, Laurence J.: Explosive Seam Welding Application to Reactor Repair. Presented at the Third Annual Conference of the Canadian Nuclear Society, Toronto, Canada, June 9, 1982.

9. Tool and Process for Explosive Joining of Tubes. Patent 4,708,280, November 24, 1987.

10. Explosive Joining, Industrial Productivity and Manufacturing Technology. NASA Spinoff, page 120, 1989.

TABLE I.-CROSS-SECTIONAL DIMENSIONS OF LINEAR RIBBON EXPLOSIVES

Explosive load, grains/ft	Explosive load, gm/m	Thickness, cm	Thickness, inch	Width, cm	Width, inch
7	.138	.051	.020	.559	.220
10	.198	.051	.020	.760	.300
15	.296	.064	.025	.800	.315
20	.395	.076	.030	.927	.365
25	.494	.089	.035	.940	.370
30	.593	.089	.035	1.295	.510

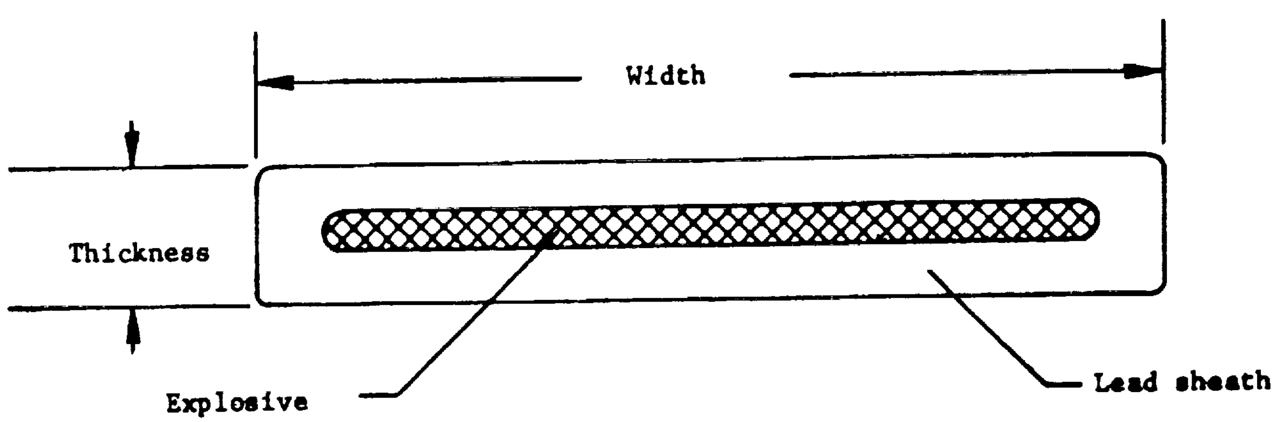


TABLE II.- LIKE METALS DEMONSTRATED JOINABLE BY EXPLOSIVE SEAM WELDING (100% STRENGTH JOINTS)

Metal	Range of Thickness cm	inch
a. Iron/steel Low-carbon, 300/400 ss	.003 - .127	.001 - .050
b. Aluminum - any fully annealed alloy and all age and work-hardened alloys, except 2024 and 6061.	.025 - .478	.010 - .188
c. Copper/brass	.025 - 0.381	.010 -.150
d. Titanium (Ti-6Al-4V)	.013 - .127	.005 -.050
e. Tantalum	.227	.090
f. Zirconium	.160	.063
g. Columbium	.081	.032

TABLE III.- METAL COMBINATIONS DEMONSTRATED JOINABLE BY EXPLOSIVE SEAM WELDING

a. Low-carbon to series 300 and 400 stainless steel in any combination.

b. All aluminum alloys and conditions are joinable to any other alloy and condition, except a combination of 2024-T3, T4, etc. to 7075-T3, T6, etc.

c. Any combination of copper, aluminum, and brass

d. Tellurium to niobium

e. Nickel to steel

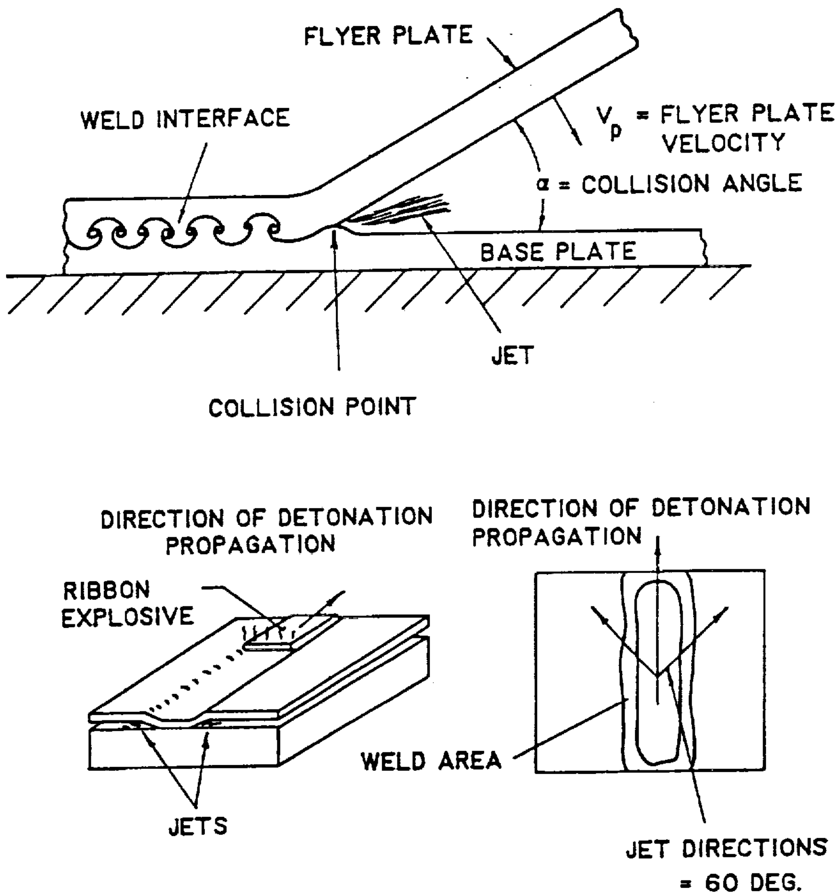


Figure 1. High-velocity, angular impacts of the two explosive joining processes: cladding in the top sketch and the NASA explosive seam welding in the bottom sketch.

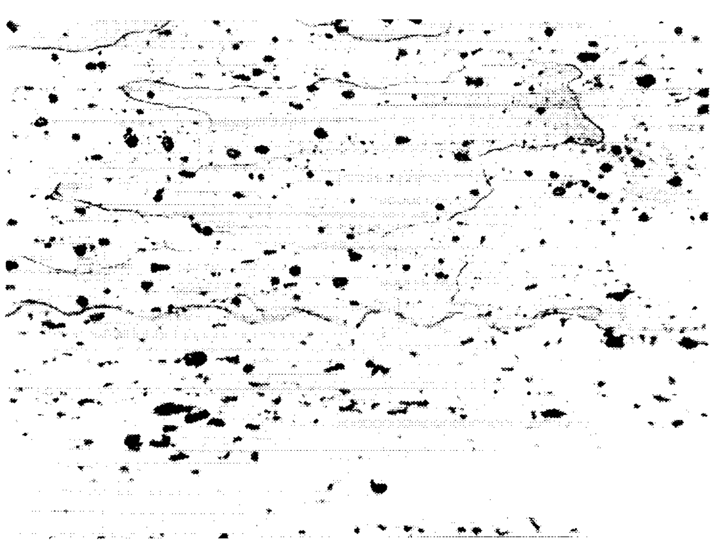


Figure 2. Microphotograph example of an explosively joined interface of 2024-T4 (top half of photograph) to 6061-T6 (lower half).

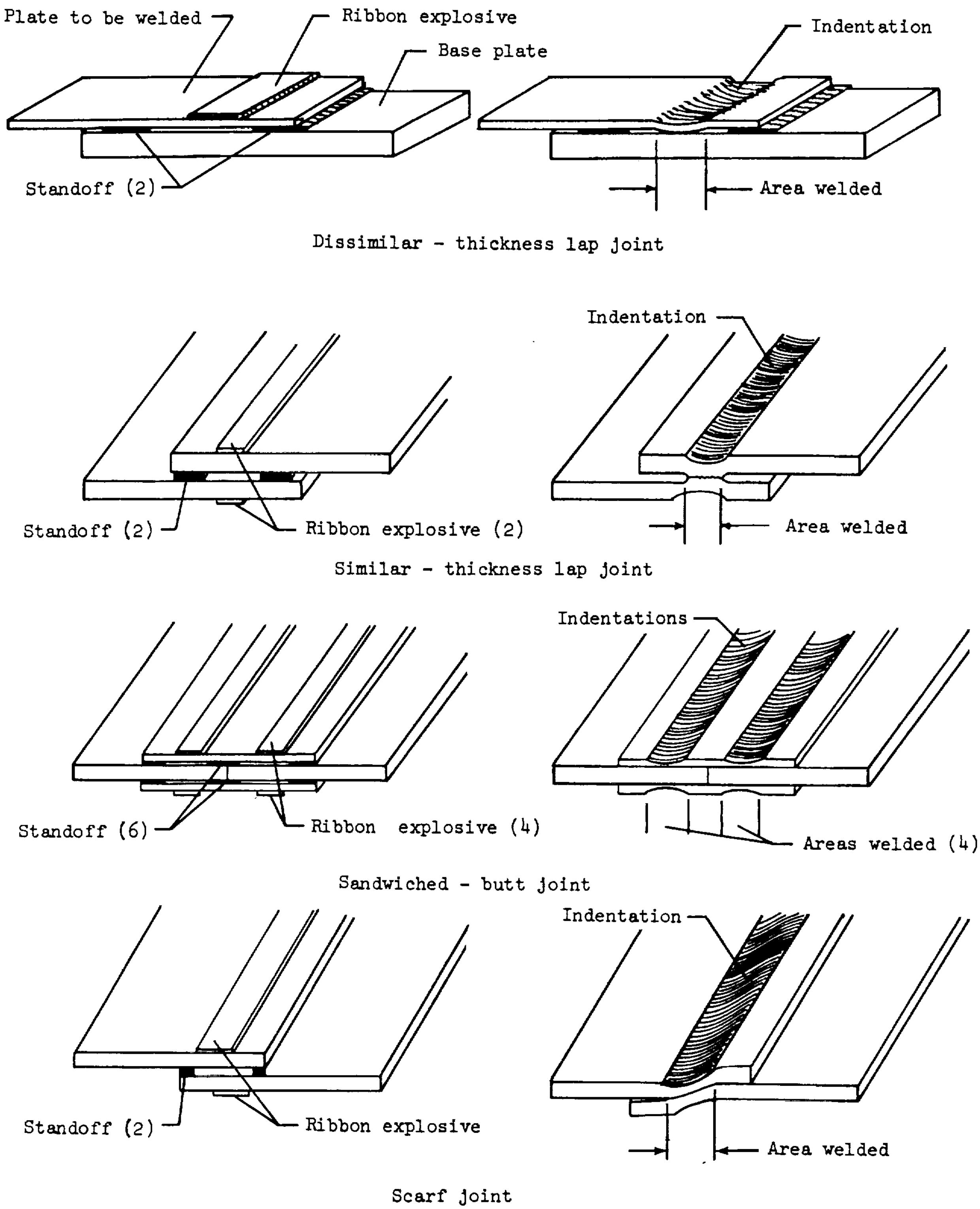


Figure 3. NASA's small-scale explosive seam welded joints.

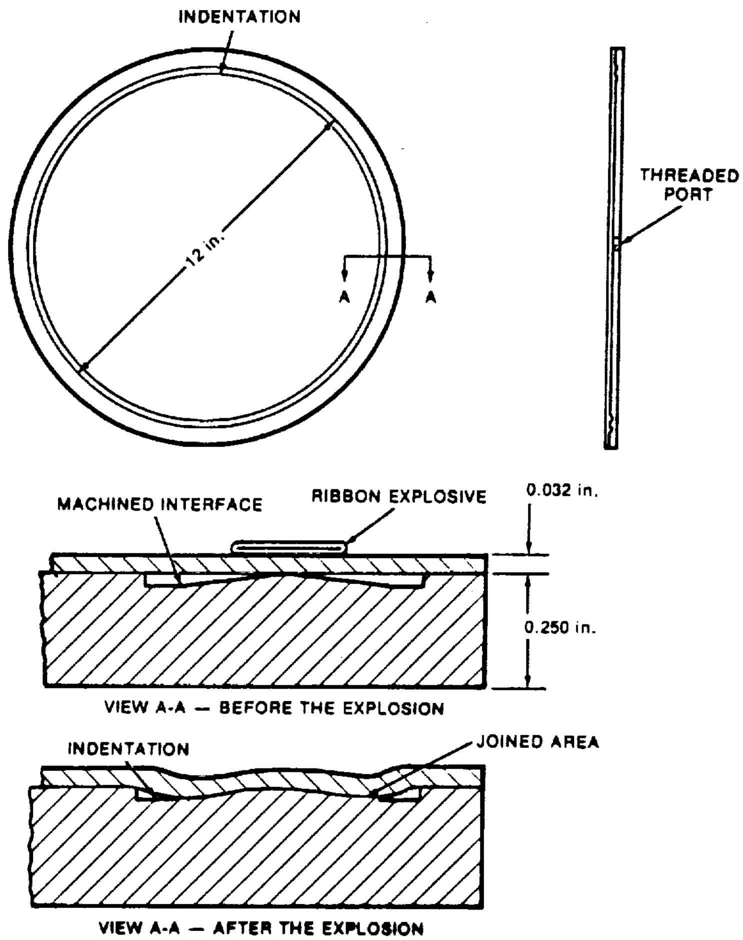


Figure 4. Approach for closing and hermetically sealing a vessel.

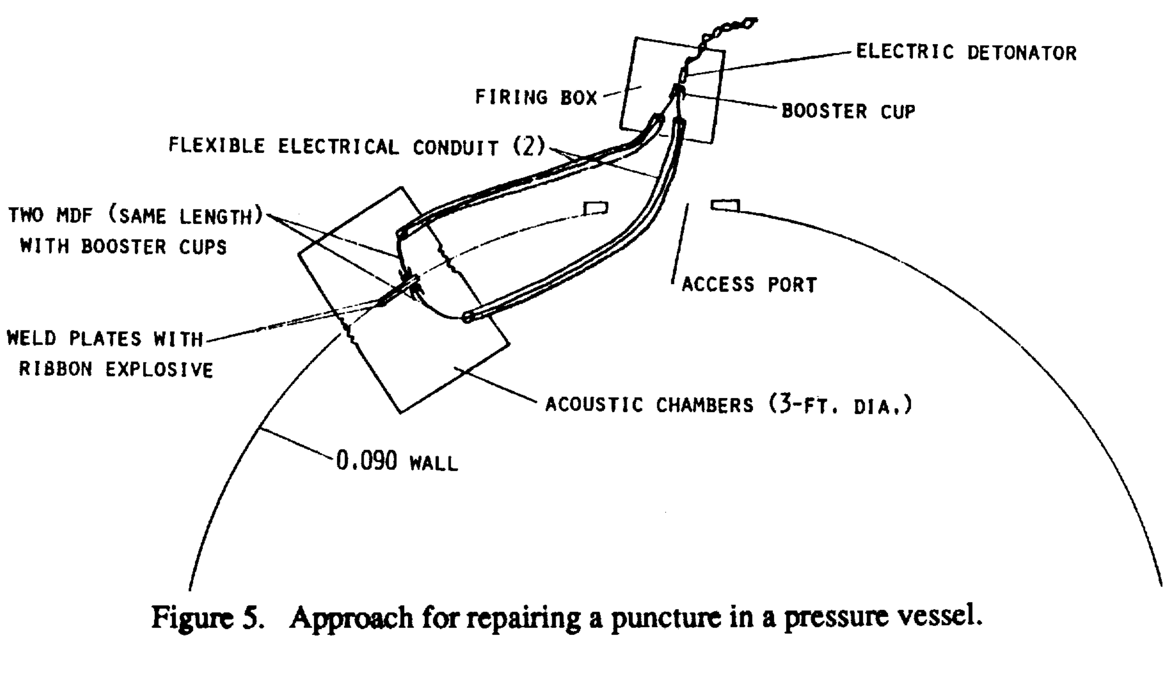


Figure 5. Approach for repairing a puncture in a pressure vessel.

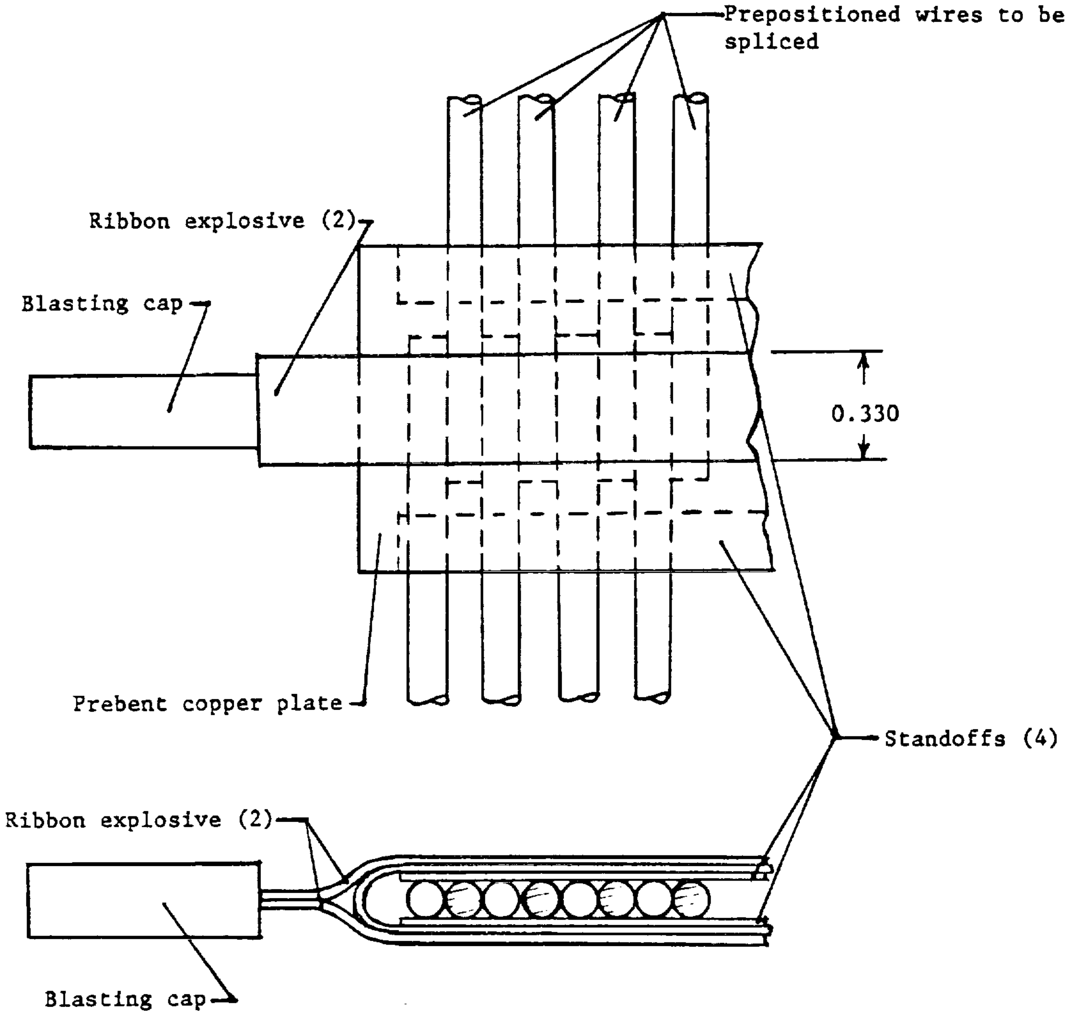


Figure 6. Explosive joining setup for splicing wires.

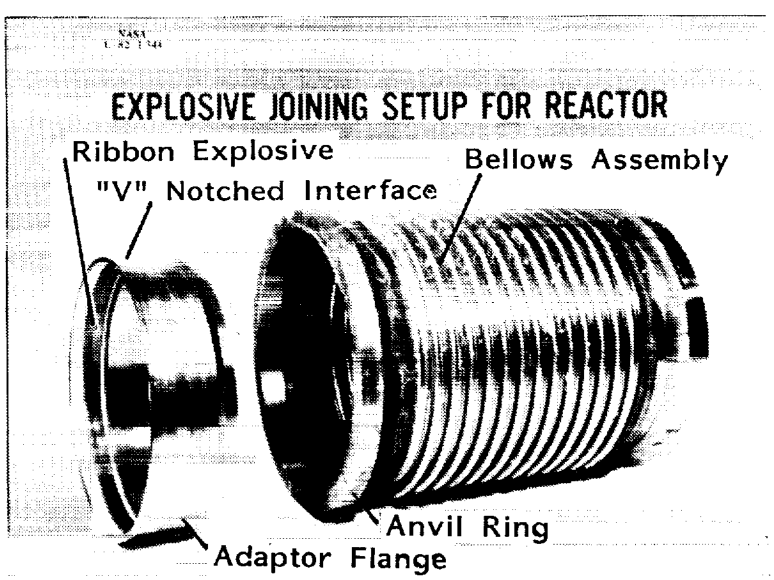


Figure 7. Setup for explosive joining adaptor flange to bellows assembly for Candu reactor.

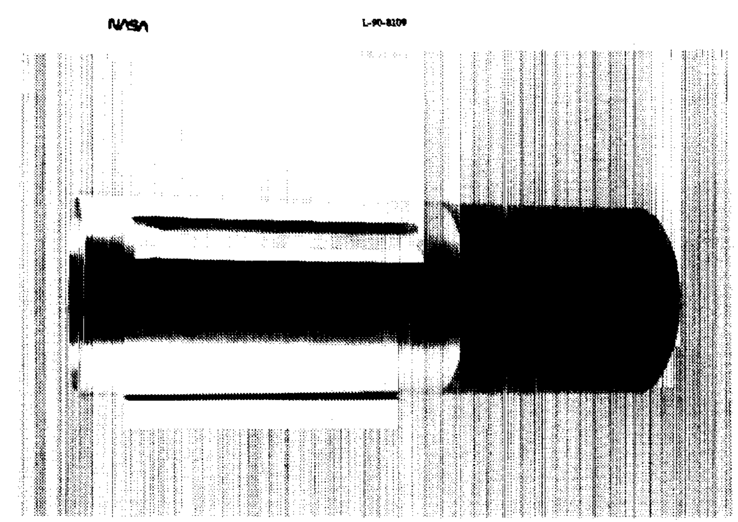


Figure 8. Cooling fins explosively bonded to Candu pressure tube.

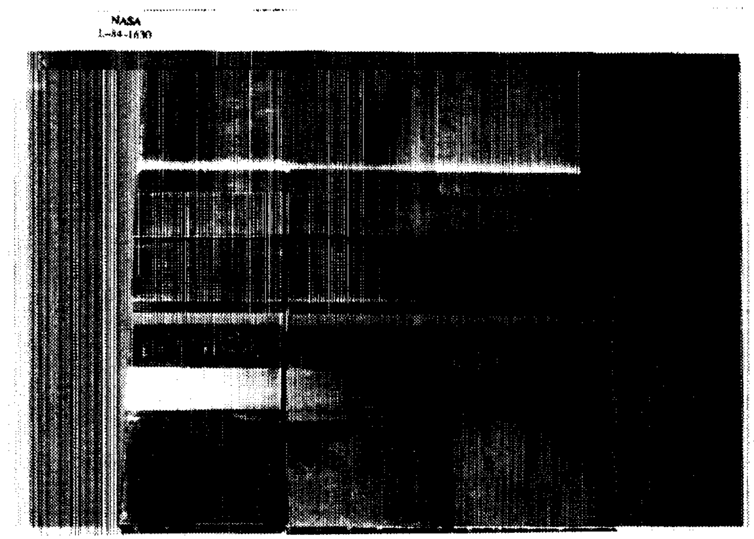


Figure 9. Preliminary evaluation of joining 316L sheet stock to (left to right) Inconel 625, Incoloy 903 and Haynes 188 with a V-notched interface.

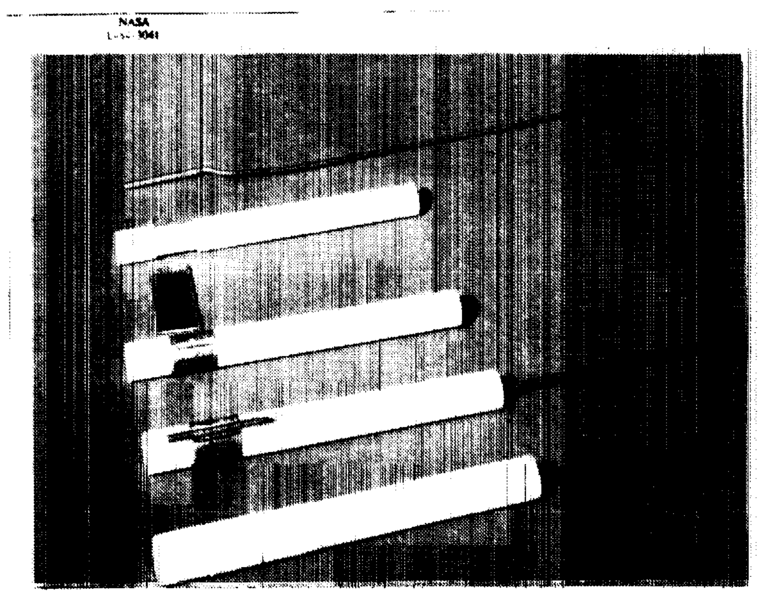


Figure 10. Assembly of the tool used for joining of tubes to fittings.

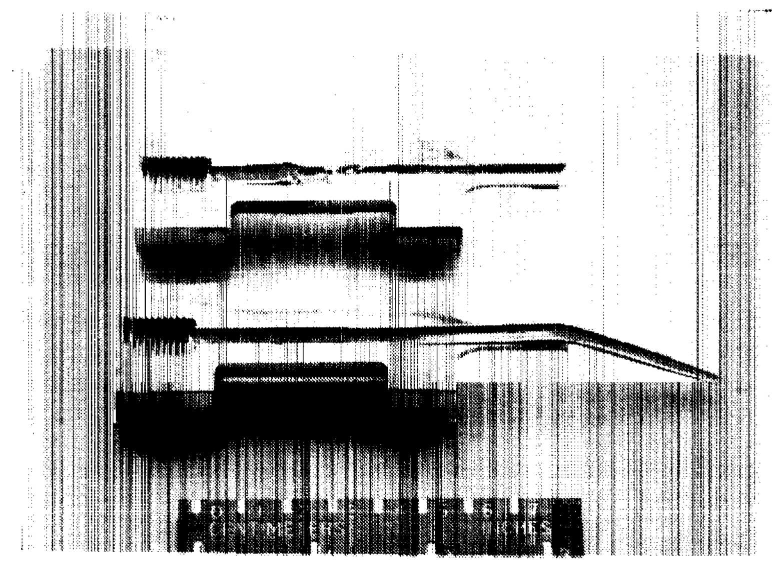


Figure 11. Explosively welded, 0.260 OD, 0.026-inch wall, 316L tube to Haynes 188 fitting with a V-notched interface. Top figure shows attempt to peel weld.

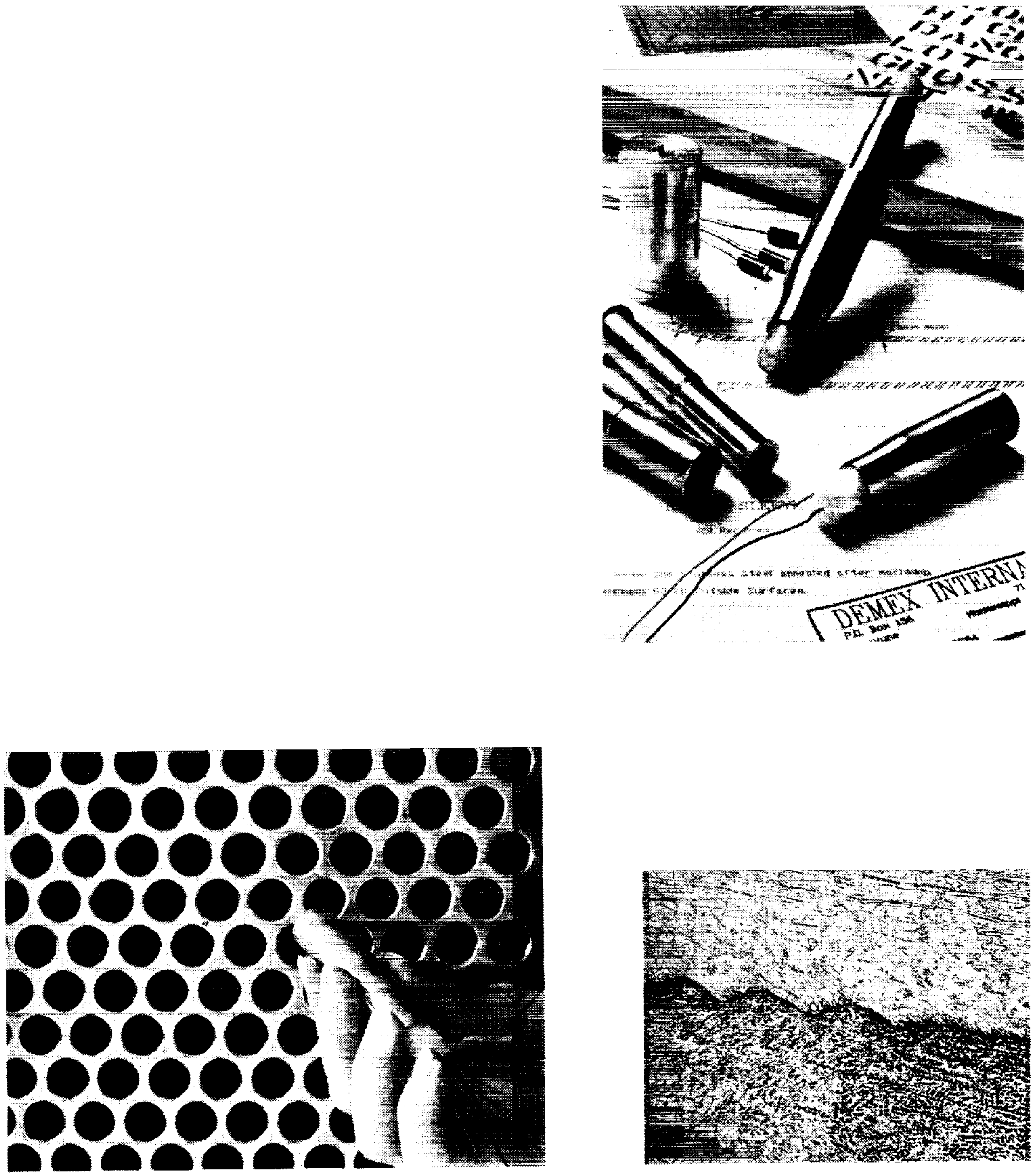


Figure 12. Tools for explosive plugging of heat exchangers Metallurgical bond is shown in the lower right.

N92-22426

Welding Technology Transfer Task/Laser Based Weld Joint Tracking System for Compressor Girth Welds

Alan Looney
NASA/Marshall Space Flight Center
MSFC, AL 35812

ABSTRACT

Sensors to control and monitor the welding operations are currently being developed at Marshall Space Flight Center. The laser based weld bead profiler/torch rotation sensor has been modified to provide a weld joint tracking system for compressor girth welds. The tracking system features a precision laser based vision sensor, automated two axis machine motion, and an industrial PC controller. The system benefits are elimination of weld repairs caused by joint tracking errors which reduces manufacturing costs and increases production output, simplification of tooling, and free costly manufacturing floor space.

INTRODUCTION

This task is a result of the technology transfer program initiated by NASA to transfer aerospace developed technology into the private sector to improve quality and productivity.

The Marshall Automated Welding System Development Program for the shuttle external tank, advanced solid rocket motor casings, and National Launch System(NLS) vehicle includes the development of optical sensors to control and monitor the welding operations. The laser based weld bead profiler/torch rotation sensor has been modified and new software developed to provide a laser based weld joint tracking system for compressor girth welds, which utilize the Gas Metal Arc Welding Process(GMAW), for Copeland Corporation.

REVIEW OF COPELAND'S WELDING PROCESS AND TOOLING

General Information

The compressor shells are .120" thick(nominal) draw quality 1008/1010 mild steel. Copeland utilizes the GMAW weld process. Copeland has five weld stations.

A typical weld cycle is listed below:

1) An operator loads a compressor shell onto the rotary table and actuates overhead air ram to hold the unit in place.
2) The operator then actuates the mechanical weld seam finding arm. After the seam is mechanically "found", a secondary arm with a hardened tracing stylus and the GMAW torch tooling contacts the lip of the lower half of the compressor shell. The table then begins to rotate and the weld portion of the cycle is initiated.
3) When the table completes one revolution plus approximately 3/4" for weld overlap, the table returns to its initial preset start point.
4) The torch tooling retracts and the air ram releases and the operator pushes the welded unit onto a section of roller conveyor ready to load another unit in place.
5) The welders average 550 to 650 weld cycles/shift.

Problem Statement

The current mechanical positioning devices are not capable of "real time" corrections due to variations in ideal weld joints of the compressor shells. Also, the current process requires special tooling(elliptical shaped gears) which offsets the shape of the part for the welding operation.

OBJECTIVE

Provide a weld joint tracking system that maintains the correct weld joint path and standoff distance which will eliminate weld defects/repairs caused by joint tracking errors during compressor girth welding operations. Demonstrate that a direct drive turntable may be used with this system which would eliminate/reduce some of the tooling and maintenance costs associated with the present tooling and reduce the floor space needed for a weld station.

SCOPE OF WORK

Modify the laser based weld bead profiler/torch rotation sensor and develop the software to provide a computer controlled weld joint tracking system which maintains the correct weld path and standoff (torch to work distance) during compressor girth welding operations. The system must be adaptable to both the existing Copeland tooling and a simple direct drive turntable setup.

Design and fabricate the fixturing for the weld joint tracking system to be compatible with the existing Copeland weld station equipment.

Perform the initial testing and demonstration of the tracking system utilizing the weld system and tooling at MSFC.

Install, checkout, and demonstrate the system at Copeland Corporation.

LASER BASED WELD TRACKING SYSTEM

System Features

The tracking system features are as follows:

1) Precision laser based vision sensor - The sensor illuminates a line across the weld joint with a pulsed, fan-shaped beam of light from a laser diode. Light reflected from the illuminated area is imaged in a camera, the shutter of which can be opened during times as short as 100ns. The laser pulse is synchronized with the opening of the shutter to maximize the amount of laser light integrated. The amount of arc light integrated is minimized as a result of keeping the opening time short. The sensor operates in conjunction with a video digitizer and a computer. By use of a geometric transformation based on the position and orientation of the camera with respect to the fan of light and the workpiece, the computer controls the position of the torch with respect to the weld joint.[1]

2) Automated two axis machine motion

3) Industrial PC controller

4) The system is invariant to travel rate changes.

Photographs of the system hardware features at MSFC are attached.

[1] NASA Tech Briefs, April 1991, Page 40

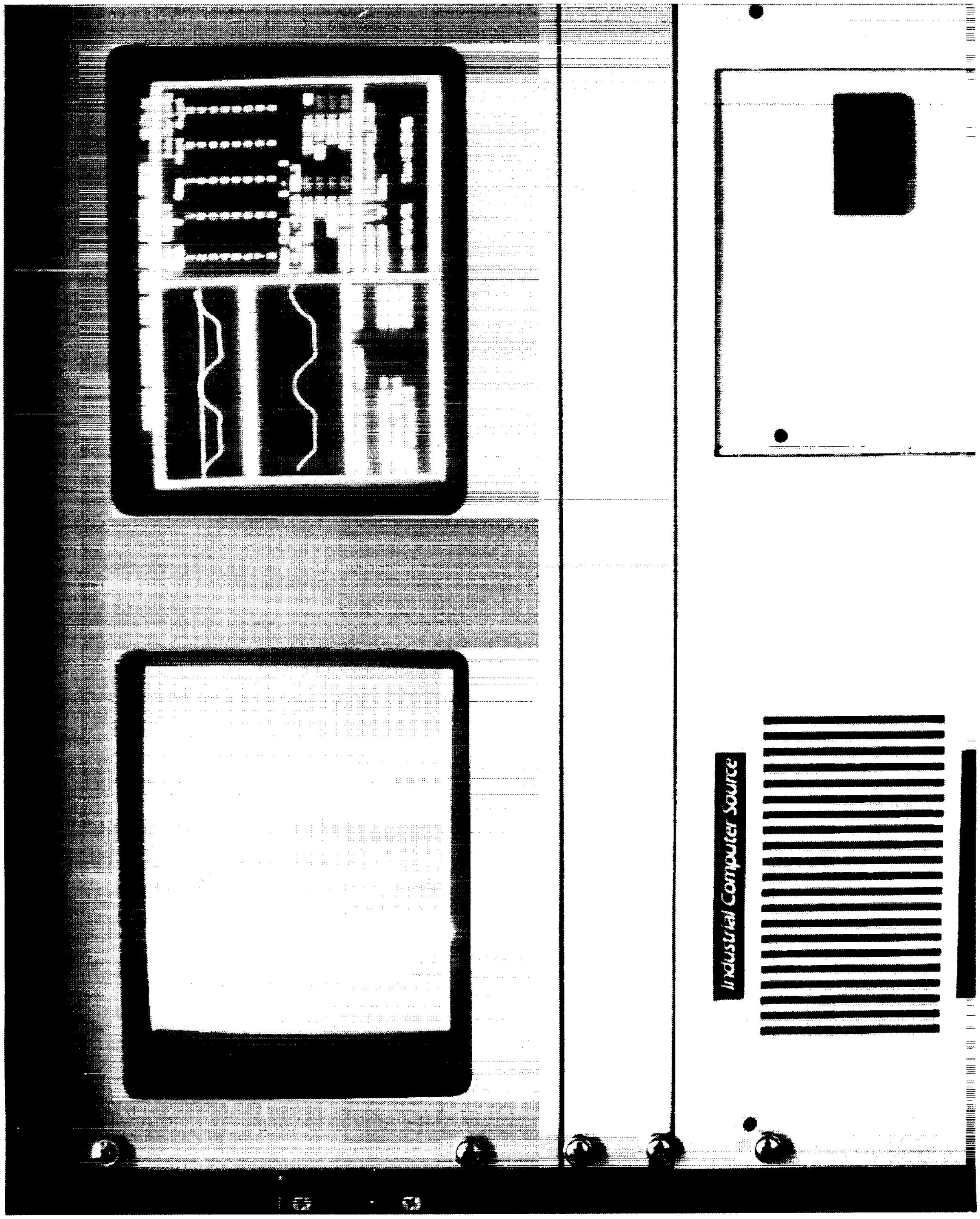
System Benefits

As a result of the system operation capability demonstrated at MSFC, the immediate benefits which have been identified are as follows:

1) Elimination of weld repairs caused by joint tracking errors which results in reduced manufacturing costs and increased production output

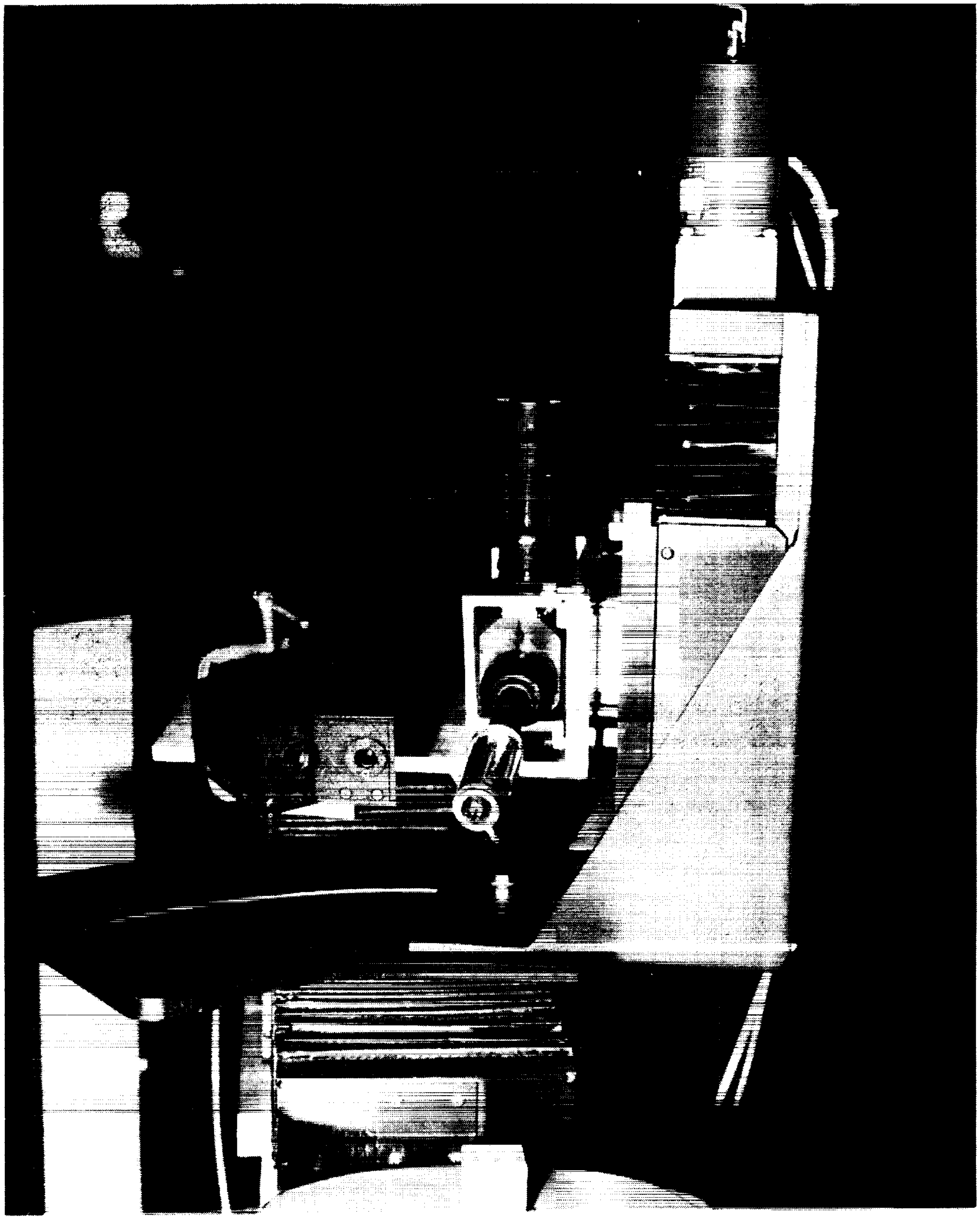
2) Simplification of tooling, i.e. the capability to use simple direct drive turntables

3) Free costly manufacturing floor space



WELD JOINT TRACKING SYSTEM AND TOOLING FOR THE WELDING TECHNOLOGY TRANSFER TASK FOR THE COPELAND CORP.

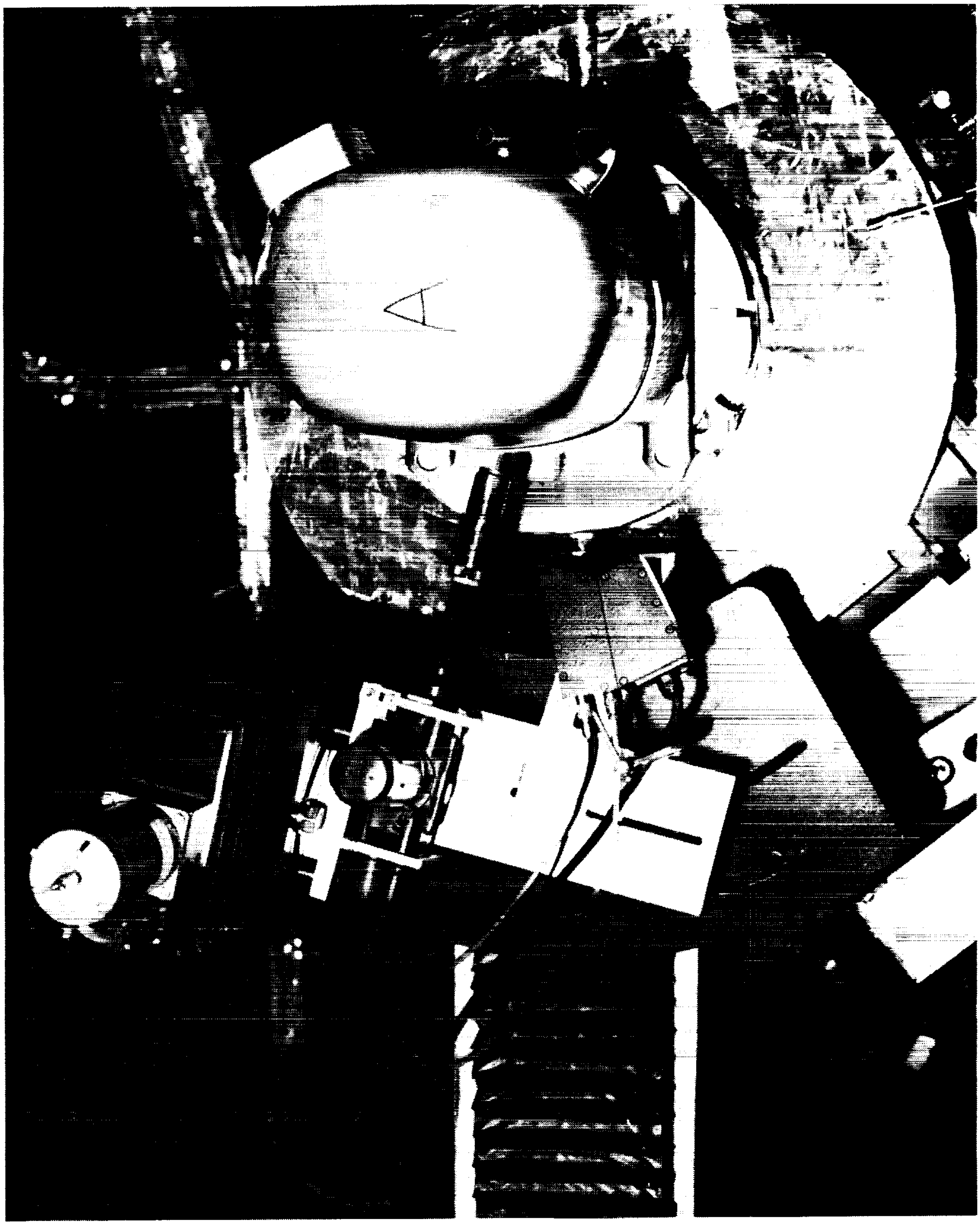
ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH



WELD JOINT TRACKING SYSTEM AND TOOLING FOR THE WELDING TECHNOLOGY TRANSFER TASK FOR THE COPELAND CORP.

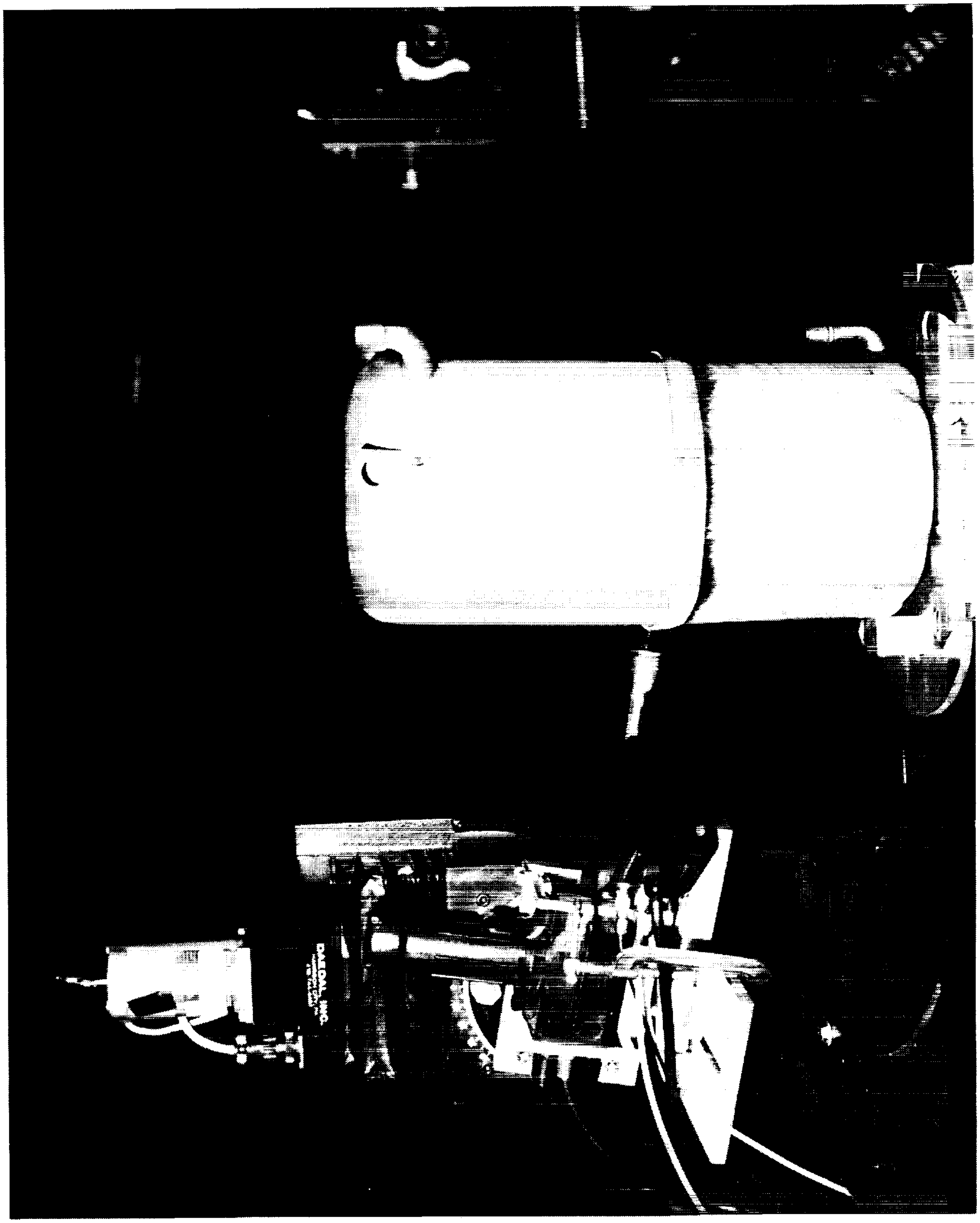
ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

WELD JOINT TRACKING SYSTEM AND TOOLING FOR THE WELDING TECHNOLOGY TRANSFER TASK FOR THE COPELAND CORP.



WELD JOINT TRACKING SYSTEM AND TOOLING FOR THE WELDING TECHNOLOGY TRANSFER TASK FOR THE COPELAND CORP.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH



WELD JOINT TRACKING SYSTEM AND TOOLING FOR THE WELDING TECHNOLOGY TRANSFER TASK FOR THE COPELAND CORP.

N92-22427

PRECISION JOINING CENTER

John W. Powell
EG&G Rocky Flats
Golden, Colorado 80402-0464

ABSTRACT

EG&G Rocky Flats and the American Welding Society (AWS) are proposing to establish a Precision Joining Center (PJC). The PJC will be a cooperatively operated center with participation from EG&G Rocky Flats, AWS, the Department of Energy (DOE) Nuclear Weapons Complex (NWC), the Colorado School of Mines, and U.S. private industry. The PJC's primary mission will be as a training center for advanced joining technologies. This will accomplish several objectives: 1) It will provide an effective mechanism to transfer joining technology from the NWC to private industry; 2) It will provide a center for testing new joining processes for the NWC and private industry ; 3) It will provide highly trained personnel to support advanced joining processes for the NWC and private industry.

INTRODUCTION

A study began at the Rocky Flats Plant in the fall of 1989 to evaluate forming a U.S. Center of Excellence (COE) in the field of joining technology. To accomplish this, the study evaluated U.S. and European welding institutes for operating structures and research and development emphasis. In addition, other Department of Energy (DOE) technology transfer programs were evaluated and noted for their successes and failures. The area of need that has emerged from this study was that of training in precision joining processes. From this has come the concept of a Precision Joining Center, with its primary mission to be a training center for high level technicians (i.e. technologists) in the use of precision joining processes.

The American Welding Society has concurred with this material need and has joined in the work of establishing this center. The Precision Joining Center will be a cooperative effort between EG&G Rocky Flats, the American Welding Society, the DOE Weapons Complex, and U.S. private industry. The center will, by the means of training, accomplish several objectives: 1) It will provide an effective means of technology transfer in precision joining methods from the DOE complex to U.S. private industry, 2) In the reverse case, the center will provide a testing ground for new technology for DOE complex modernization; 3) The center will fill the education gap between existing vocational schools and engineering colleges; 4) The center will be a resource for small to medium sized industrial companies for precision joining processes.

PROPOSED STRUCTURE

The structuring of the Precision Joining Center has several important considerations.

Curriculum

The first consideration is the curriculum. The curriculum will be established by industrial input to keep it current with the on-going needs of industry. The initial curriculum outlined in this article is a result of an industrial workshop held for the purpose of its development. The established curriculum will be continually updated to meet new needs through customer contacts.

Course Structure

For the center to operate successfully with industry, the time and financial impact on participating companies must be minimized. This will be accomplished by making the core courses of the program portable so that they can be taught at any location. The class starting times will be flexible. This will reduce travel expenses and time away from the job for the students participating in the program.

The hands-on training will be accomplished at the center and will require travel by the participating student.

Equipment

For an advanced joining center to be effective, it must be equipped with state-of- the- art equipment. To accomplish this, manufacturers will be asked to participate by consignment of advanced joining equipment. Then, as equipment needs to be updated, the older machines will be returned to the suppliers as the new models are installed. The advantage for the suppliers will be first hand experience on their machines by potential future customers.

Staffing

Staffing for the center will come from the U.S. joining community based on minimum qualifications. However, staffing for the Precision Joining Center facility will be assured by EG&G Rocky Flats, and the staffing for core courses taught off-site will be assured by the American Welding Society.

Scope of Training

The scope of training will be interdisciplinary and will include the following subject matter:

- Joining Technology
- Vacuum Technology
- Power Systems
- Data Acquisition
- Servo-Mechanisms and Robotics
- Computer Controls
- Record Keeping and Palentability
- Weld Defects and Inspection
- Metallurgical Considerations
- Heat Flow Considerations
- Fixtures and Tools
- Metrology
- Health, Safety, and Environment

The Welding Technologist

The curriculum will be designed to train hands-on technologists. These are the people that companies presently develop over several years of on-the-job experiences. They are the people who understand equipment, processes, and how to get projects done. This program will be designed to create these individuals faster, with fewer deficiencies in their training, and with a broader range of joining knowledge and experience.

Entrance Requirements

With the above objectives in mind, there will be minimum requirements for entrance into the program. These requirements are as follows:

- A graduate of a trade school, community college or equivalent.
- Or a minimum of three (3) years of technical experience.
- Plus recommendations

Graduate Profile

The training program objective will be to produce graduates with the following profile:

- Proven ability to:
  - Operate advanced welding/joining equipment
  - Communicate effectively with engineers, welders, and managers
  - Understand and direct welding and support processes
  - Perform joining R&D with minimal oversight
- Documentation Skills
- Analytical Skills
- Measurement Skills

The Curriculum

The curriculum will be divided into two major segments. The first segment will be composed of eight one-week core courses. These courses will be taught at off-site locations and at the Precision Joining Center. It is anticipated that most students would attend the core courses at a location in or near to their normal workplace. The second segment will be composed of three two-week courses taught at the Center. These courses will include considerable hands-on experience with state-of-the-art joining equipment.

Core Courses

The core courses will be as follows:

1. Precision Process Controls and Metrology

2. Welding Process Selection

3. Elements of Welding Metallurgy

4a. Fusion Welding Parameters and Fusion Zone Profiles

4b. Solid State Bonding and Brazing Processes

5. Weld Properties, Design, and Defects

6. Weld Testing and Defect Detection

7. Weld Parameter Development and Statistical Process Control

8. Welding Problem Analysis and Technical Communications

Joining Process Courses

These courses are hands-on courses designed to give the student practical experience with process control and design. The process course will be subdivided into three categories: arc processes, beam processes, and non-fusion processes. Each of these categories will have three two-week sessions of classroom and laboratory work. These hands-on sessions will cover a variety of topics including the following:

- Process variables and Controls
- Interactive Control Systems
- Computerized Joining Systems
- Data Acquisition
- Support Systems
    - Basic Electronics
    - Vacuum Systems
    - Power Systems
    - Servo Mechanisms
- Tooling and Fixturing
- Application of QA Concepts

Certification

As a student progresses through the program permanent records will be maintained. In addition, a Continuing Educational Unit will be given for each successfully completed course. After completion of the program, the student will receive a certification of completion. His records will be accessible through the Precision Joining Center registrar.

SUMMARY

The establishment of the Precision Joining Center will meet several current U.S. industrial needs. It will provide a mechanism to transfer joining technology between the NWC and private industry. It will help meet the need for trained joining technologists to operate industrial precision joining processes. And, it will provide a resource for small and medium size companies in advanced joining systems.

BIOTECHNOLOGY

(Session A2/Room C1)

Tuesday December 3, 1991

- Cooperative Research and Development Opportunities with the National Cancer Institute
- Technologies for the Marketplace from the Centers for Disease Control
- Enhancement of Biological Control Agents for Use Against Forest Insect Pests and Diseases
- Use of T7 Polymerase to Direct Expression of Outer Surface Protein A (OspA) from the Lyme Disease Spriochete, *Borrelia burgdorferi*

BIOTECHNOLOGY

N92-22428

COOPERATIVE RESEARCH AND DEVELOPMENT OPPORTUNITIES WITH THE NATIONAL CANCER INSTITUTE

Kathleen Sybert, Ph.D.
Deputy Director
Office of Technology Development
National Cancer Institute
Bethesda, Maryland 20892

ABSTRACT

The Office of Technology Development (OTD) of the National Cancer Institute (NCI) is responsible for negotiating Cooperative Research and Development Agreements (CRADAs), whereby the knowledge resulting from NCI investigators' government-sponsored research is developed in collaboration with universities and/or industry into new products of importance for the diagnosis and treatment of cancer and acquired immunodeficiency syndrome (AIDS). The NCI has recently executed a unique "clinical trials" CRADA and is developing a model agreement based upon it for the development and commercialization of products for the diagnosis and treatment of cancer and AIDS. NCI drug screening, preclinical testing, clinical trials, and AIDS program capabilities form the basis for this new technology development/technology transfer vehicle. NCI's extensive drug screening program and "designer foods" program serve as potential sources of investigational new drugs (INDs) and cancer preventatives. Collaborations between NCI and pharmaceutical companies having the facilities, experience, and expertise necessary to develop INDs into approved drugs available to the public are being encouraged where the companies have proprietary rights to INDs, or where NCI has proprietary rights to INDs and invites companies to respond to a collaborator announcement published in the Federal Register. The joint efforts of the NCI and the chosen collaborator are designed to generate the data necessary to obtain pharmaceutic regulatory approval from the Food and Drug Administration (FDA) to market the drugs developed, and thereby make them available to health care providers for the diagnosis and treatment of cancer and AIDS.

INTRODUCTION

The Office of Technology Development is organizationally located in the Office of the Director of the National Cancer Institute. The National Cancer Institute is one of the thirteen Institutes of the National Institutes of Health. The National Institutes of Health is a part of the Public Health Service, which in turn is part of the Department of Health and Human Services.

The mission of the National Institutes of Health is to conduct biomedical and behavioral research into the treatment and control of disease, to positively benefit the health of the American people. As a main source of funding for medical research in the United States, the National Institutes of Health is a national resource for the evaluation of new disease therapies. Historically, the National Cancer Institute has been the most important effector in the discovery and development of new anticancer agents. More recently, the National Cancer Institute has also played a leading role in the discovery and development of anti-AIDS agents such as azidothymidine (AZT) and dideoxyinosine (ddI).

Cancer and AIDS research efforts at the National Cancer Institute are both basic and applied. Basic biomedical and molecular biology research being conducted include sequencing studies and studies of the molecular actions of drugs and mutagens. Applied diagnostic and treatment research being conducted includes: methods and materials to detect early cancer; methods and materials to detect human immunodeficiency virus (HIV) infection; and clinical trials to determine the efficacy of anti-cancer and anti-AIDS agents.

The Office of Technology Development serves as the National Cancer Institute's focal point for the implementation of the laws, policies, rules and regulations related to the implementation of the Federal Technology Transfer Act of 1986. The Office is responsible for the administration of activities related to collaborative agreements,

 PRECEDING PAGE BLANK NOT FILMED

confidentiality agreements, material transfer agreements, inventions, patents, licensing and royalties. Office of Technology Development staff members provide advice, guidance and assistance to National Cancer Institute scientists and staff, and review and analyze planned agreements to ensure that they comport with applicable National Cancer Institute, National Institutes of Health, Public Health Service, and Department of Health and Human Services policy and procedures.

The Office of Technology Development interacts with several other National Institutes of Health technology transfer components. Among these are: the Patent Policy Board; the National Institutes of Health Office of Technology Transfer; the various Institute Technology Development Coordinators; and extramural components, including the Office of Extramural Programs. The Patent Policy Board oversees the technology transfer program for the Public Health Service, and makes policy recommendations to Public Health Service agency heads. The Office of Technology Transfer serves the following centralized functions for the National Institutes of Health: preparation, filing and prosecution of domestic and foreign patent applications, utilizing the services of its own Patent Branch, outside contract attorneys, and the National Technical Information Service; marketing and licensing of technology and inventions, either directly or through the Center for the Utilization of Federal Technology, Department of Commerce; and drafting of model agreements. Each of the Institutes of the National Institutes of Health has a Technology Development Coordinator who participates in the interactive stages of patent prosecution and licensing activities. Two Institutes also have established offices for the transfer of technology. The National Cancer Institute has established the Office of Technology Development; and the National Institute of Allergy and Infectious Diseases has established the Technology Transfer Branch. The extramural component includes grants, contracts, cooperative research and development agreements, informal clinical trial agreements, memoranda of understanding, and confidentiality agreements; for drug screening and discovery, and clinical trials with private industry. Through the office of Extramural Programs, grantee institutions report patentable inventions developed using Federal funds, and communicate the grantees' decisions regarding the assignment of rights to those inventions.

The Office of Technology Development was established as the result of a series of legislative enactments that were designed to transfer Federal technology to industry, to state and local governments, and to universities. Among these legislative enactments are: Federal patent law; the Bayh-Dole Patent and Trademark Act of 1980; the Stevenson-Wydler Technology Innovation Act of 1980; the Federal Technology Transfer Act of 1986; and Executive Order 125991 of 1987. Federal patent law (35 U. S. C. at §§ 200-212) authorizes the licensing of Government-owned patent rights. Under it, "... Each Federal agency is authorized to ... grant non-exclusive, exclusive, or partially exclusive licenses under federally owned patent applications, patents, or other forms of protection obtained, royalty-free or for royalties or other consideration, and on such terms and conditions, including the grant of a license of the rights of enforcement pursuant to the provisions of chapter 29 of this title as determined appropriate in the public interest; ... ." The Bayh-Dole Act of 1980 allows nonprofit organizations and small business to retain rights to inventions (patents) or other intellectual property (trademarks, copyrights) developed under Federal grant or contract funding. This Act has been amended to allow Federal laboratories operated by nonprofit organizations to similarly retain intellectual property rights and commercialize Federally sponsored inventions. The Stevenson-Wydler Technology Innovation Act of 1980 establishes the policy that it is the duty of each Federal laboratory to transfer Federal technology to industry, to state and local governments, and to universities. However, it was not until the passage of the Technology Transfer of 1986, that Federal laboratories had effective incentives to encourage both the Federal scientists and collaborators in the state, local and private sectors to participate and accomplish this goal.

The Federal Technology Transfer Act (15 U.S.C. § 3710, "Utilization of Federal Technology") of 1986 amended the Stevenson-Wydler Technology Innovation Act of 1980 to authorized Federal laboratories to enter into Cooperative Research and Development Agreements (CRADAs), and to grant intellectual property rights in advance to collaborators for inventions made in whole or in part by Federal employees under the CRADA. The Federal Technology Transfer Act requires that agencies establish cash award programs for inventors and non-inventors for their contributions to technology transfer, and that laboratory directors recognize the competitive advantage Congress intended technology transfer to grant United States business. A 1989 amendment to the Federal Technology Transfer Act authorizes government-owned, contractor-operated (GOCO) facilities to enter into CRADAs so that technology developed for the government by private parties can also be transferred. Executive Order 12591 of April 10, 1987, "Facilitating Access to Science and Technology", orders Federal laboratories to transfer new knowledge from the research laboratories to universities and to the private sector (to "privatize" Federal research inventions), assisting them in the development of new products and processes, thereby broadening our national technology base, and strengthening United States manufacturers' competitive position in the international economic arena.

Legislation is pending that would further enable technology transfer from the Federal sector. Currently, Federal employees are prohibited from copyrighting any work developed as part of their official duties. Under consideration is the protection by copyright of computer software and other material developed by Federal employees under a CRADA.

MODEL COOPERATIVE RESEARCH AND DEVELOPMENT AGREEMENT AND RELATED POLICY

The National Institutes of Health/Alcohol, Drug Abuse and Mental Health Administration (NIH/ADAMHA) Model Cooperative Research and Development Agreement is a contractual mechanism for technology transfer, under which information and materials are exchanged between collaborators. The Articles and Appendices of the NIH/ADAMHA Model Agreement are as follows, with National Cancer Institute additions indicated:

Article 1.	Introduction
Article 2.	Definitions
Article 3.	Cooperative Research
Article 4.	Reports
Article 5.	Financial and Staffing Obligations
Article 6.	Title to Property
Article 7.	Intellectual Property Rights and Applications
Article 8.	Licensing
Article 9.	Proprietary Rights and Publication*
Article 10.	Representations and Warranties**
Article 11.	Termination
Article 12.	Disputes
Article 13.	Liability
Article 14.	Miscellaneous
Article 15.	Duration of Agreement
Appendix A	NIH/ADAMHA Policy Statement on CRADAs and Intellectual Property Licensing
Appendix B	Research Plan
Appendix C	Financial and Staffing Contributions of the Parties
Appendix D	Exceptions or Modifications to this CRADA
*	NCI adds "Intellectual Contributions of the Parties" to this Article
**	NCI adds "Potential Patentability of Subject Inventions" to this Article

Several NIH/ADAMHA policy considerations impact upon the terms agreed to by the Government under a CRADA. National Cancer Institute investigators are free to choose their research topics, provided their choices are consistent with the mission of the Institute and the program of their particular laboratories. This research freedom cannot be constrained by the conditions of a CRADA or a related licensing agreement. Nor can the scientists' participation in a cooperative research plan restrict his ability to disseminate research results freely in both publications and public fora. Provision is made for reasonable delays in this dissemination only for the purpose of filing patent application(s), which filing is, in fact, encouraged.

Under a CRADA there is an exchange between the collaborators of intellectual and/or technical resources that are not otherwise reasonably available. Provision is made for personnel, services, facilities and equipment to be exchanged between the collaborator and the National Cancer Institute. Funds may flow from the collaborator to the National Cancer Institute; however, in no case may funds flow from the National Cancer Institute to the collaborator.

The CRADA collaborator is chosen on the basis of scientific expertise and commercialization capabilities. When the Government has the intellectual lead, a Federal Register notification may be appropriate. NIH/ADAMHA will give special consideration to entering into CRADAs with small business firms and consortia involving small business firms. Further, preference will be given to businesses located in the United States, or which agree to manufacture substantially in the United States products which embody inventions developed under CRADAs.

Proprietary information may be exchanged and maintained as confidential under a CRADA, since Freedom of Information does not require the release of "trade secrets and commercial or financial information ... [that] are privileged and confidential". Secrecy is routinely maintained under National Cancer Institute drug screening programs and during the Food and Drug Administration regulatory approval process. Proprietary information under CRADAs is maintained as confidential as long as is necessary to accomplish the research plan. However, in no case will secrecy be maintained once an invention has been patented.

The time-limited right to exclude others from making, using, or selling an invention, which is conveyed by a license is used as an incentive for collaborators to invest in product development under a CRADA. Time-limited options to negotiate non-exclusive, partially exclusive, or exclusive licenses may be granted in advance to CRADA collaborators. When granting licenses to inventions developed wholly by NIH/ADAMHA investigators or jointly with a collaborator under a CRADA, the Government retains a nonexclusive, irrevocable, paid-up license to practice the invention or to have the invention practiced throughout the world by or on behalf of the U.S. Government. The Government also requires the grant of a research license for inventions made wholly by a collaborator under a CRADA. Further, NIH/ADAMHA reserve the right under any license granted to the collaborator, to grant research licenses to third parties.

Requests for exclusive commercialization licenses must include a development and commercialization plan. After an exclusive license is granted, reports on progress toward utilization of the invention are required. NIH/ADAMHA reserves the right to grant several, separate exclusive licenses in various fields of use. Exclusive licenses will have "best effort" clauses, and may be terminated when a licensee is not actively engaged in an effort to produce product, or a licensee cannot meet market demand. Exclusive licensees must not unreasonably deny requests for sublicense in unused fields of use, or requests for cross license rights from future CRADA collaborators when derivative rights are necessary for a CRADA to go forward, and the exclusive licensee has been given a reasonable opportunity to be a party to the CRADA.

A pricing clause will be found in every CRADA since policy states that "DHHS has a concern that there be a reasonable relationship between the pricing of a licensed product, the public investment in that product, and the health and safety needs of the public. Accordingly, exclusive commercialization licenses granted for NIH/ADAMHA intellectual property rights may require that this relationship be supported by reasonable evidence."

COOPERATIVE RESEARCH AND DEVELOPMENT AGREEMENTS FOR THE CLINICAL DEVELOPMENT OF ANTI-CANCER AND/OR ANTI-AIDS AGENTS

When a compound is discovered or developed by the National Cancer Institute that shows promise as an anti-cancer or an anti-AIDS agent, the involvement of a private firm is sought, since the National Cancer Institute does not take drugs to market. The early involvement of a pharmaceutical firm permits substantial cost sharing between the Government and the private sector, and can speed the commercial availability of effective agents. Industry is asked to invest its resources to bring an agent from the discovery stage through subsequent development, clinical trials, regulatory approvals, and ultimately into commercial production. An exclusive license may be granted to the industrial collaborator in cases where substantial additional risk, time and cost must be undertaken prior to successful commercialization. An option to negotiate such a license may be granted in advance under a CRADA for the clinical development of an agent.

Under a CRADA for the clinical development of an agent, national, multicenter clinical trials in various research settings, and trials in combination with other agents, can be planned and coordinated. Efforts to investigate and evaluate alternative sources of the agent can also be planned and coordinated.

In addition to standard CRADA provisions, a CRADA for the clinical development of an anti-cancer or anti-AIDS agent may also contain terms regarding:

The formation and functioning of a Clinical Research Team, or Steering Committee

Provision for supply of drug up to New Drug Application (NDA) filing, including amounts for compassionate use

Provision for Investigational New Drug Application (IND) sponsorship, and for cross-referencing of INDs

Summary of completed and ongoing preclinical and clinical testing and data

Provision for data collection

Provision for New Drug Application (NDA) filing

Provision of support staff for Group C distribution

Conveyance of orphan drug status

License of pre-dating compounds or methods of use apart from CRADA

Publication clause: "Nothing shall prevent the timely publication of the results of clinical trials or pre-clinical research."

A recently executed example of a CRADA for Clinical Development is one for the clinical development of taxol. Taxol is a promising new drug for the treatment of refractory ovarian cancer. Taxol was discovered during the NCI-sponsored screening of extracts of over 35,000 plant species. It is a natural product that is in short supply. In return for the significant financial investment of procuring sufficient taxol for clinical trials, and for exploring alternative natural and synthetic sources, and for prosecuting the NDA, the collaborator has been given exclusive access to the clinical data, and conveyance of the orphan drug status that was granted taxol by the FDA. This is an unusual CRADA, since taxol is not patented, and other companies are, therefore, free to work on taxol, as well as its analogues.

NCI'S UNIQUE RESEARCH CAPABILITIES

NCI has research capabilities uniquely suited to cooperative research for the clinical development of anticancer and anti-AIDS agents. NCI has the largest staff and funding among the Institutes of the National Institutes of Health, and possesses unique capabilities in support of clinical development efforts. NCI's drug screening and development programs include acquisition and synthesis of novel compounds, *in vitro* screening, *in vivo* assays, preclinical testing, clinical trials, and AIDS program capabilities.

NCI's Clinical Therapy Evaluation Program (CTEP), within the Division of Cancer Treatment, is researching the uses of therapeutic compounds supplied by intramural as well as extramural researchers and pharmaceutical companies. CTEP designs and implements the development plans for new agents; it files INDs with FDA, permitting DCT to act as a drug sponsor; and it is responsible for the contracts and cooperative agreements under which most clinical testing takes place. Through grants, contracts and cooperative agreements, NCI funds a large, multicenter clinical trials network, including cooperative groups, new drug development contractors, and other investigators at cancer centers and university hospitals. More than 500 investigators and approximately 500 institutions are involved; and over 100 compounds are currently in various stages of clinical testing.

NCI oversees the only Government Owned Contractor Operated (GOCO) facility at NIH, at the Frederick Cancer Research and Development Center (NCI-FCRDC). Five separate NCI-FCRDC contract operations provide: basic research, operations and technical support, computer services, library services, and animal production. The following NCI Divisions have intramural, extramural and clinical activities at NCI-FCRDC: the Division of Cancer Etiology (DCE); the Division of Cancer Treatment (DCT, including the Developmental Therapeutics Program and Biological Response Modifiers Program); the Division of Cancer Biology, Diagnosis and Centers (DCBDC), and the Division of Cancer Prevention and Control (DCPC). GOCO contractors at NCI-FCRDC are engaged in basic research into the causes and biology of cancer, and furnish the following technical support services: support for new DCT programs for *in vitro* screening of antitumor and anti-AIDS compounds, including the development of large-scale natural products extraction capability; support for the Biological Response Modifiers Program's experimental clinical trials at a nearby outpatient facility and an inpatient capability located at a local hospital; support for AIDS vaccine development program efforts at NCI-FCRDC and an outside subcontract network; repositories for cell lines, natural product compounds, and other research materials; managing the NCI Supercomputer Center, with special emphasis on mathematical biology in biomedical research.

NCI's Cancer Network includes: Cancer Information Service (CIS), a national toll-free telephone service that provides immediate answers to cancer-related questions from cancer patients, families, the public, and health professionals; Cancer Centers, a program of cancer research centers across the country which significantly contributes to progress in basic research, clinical studies, and cancer prevention and control; Community Clinical Oncology Program (CCOP), a program affording community physicians and their patients the opportunity to participate in NCI-approved cancer treatment and cancer prevention and control clinical trials; Physicians Data Query (PDQ), an on-line computer system that provides state-of-the-art information on cancer detection, diagnosis and treatment; Cooperative Group Outreach Program (CGOP), designed to increase patient enrollment in clinical trials and to upgrade the skills of community physicians and other health professionals; Surveillance, Epidemiology, and End Results (SEER) Program, population-based cancer registries that permit monitoring of cancer incidence, mortality and survival, and is a key tool for assessing progress against cancer.

RESEARCH INITIATIVES WITHIN THE DIVISIONS OF THE NATIONAL CANCER INSTITUTE

Research initiatives within the Divisions of the National Cancer Institute of interest to potential CRADA collaborators are as follows. Research initiatives within the Division of Cancer Etiology (DCE) include dietary mutagen studies, and an investigation of the relationship between human papilloma viruses and cancer risk. Within the Division of Cancer Biology, Diagnosis and Centers (DCBDC), new discoveries are providing new strategies for potentially inhibiting cancer invasion metastasis formation and growth. Using a panel of about 60 human tumor cell lines, the Division of Cancer Treatment (DCT) is routinely screening about 20,000 synthetic compounds and natural products extracts annually for anti-cancer activity. Screening of potential anti-AIDS drugs also has been carried out in HIV-infected cells at a rate of about 20,000 synthetic compounds and natural products extracts annually. DCT's research activities also include: clinical strategies to overcome multi-drug resistance; human gene therapy; clinical development of taxol; adoptive immunotherapy; and tumor suppressor genes.

NCI PRECLINICAL AND CLINICAL TREATMENT RESEARCH EFFORTS

Pharmaceutical industry interactions with NCI for the testing and joint development of agents are possible at all stages of antitumor screening, preclinical toxicology, and clinical testing.

NCI's preclinical research efforts include new drug discovery through natural product screening and rational drug design; and the preclinical discovery and development of biological response modifiers, including such research areas as monoclonal antibodies, cell-mediated cytotoxic therapy, targeting of growth factor oncogenes and tumor-tumor suppressor genes, and hematopoietic growth factors

NCI's clinical research efforts include: enhancement of the effectiveness of chemotherapy; immunotherapy; gene therapy; differentiation therapy; radiation oncology; diagnostic imaging; and clinical trials, including investigational drug and chemoprevention trials.

RECENT/CURRENT NATIONAL CANCER INSTITUTE CRADAS

AIDS Vaccine Development: HIV gp160

Generation and Characterization of Monoclonal Antibodies to Carcinoma Associated Antigens

Retroviral-Mediated Transfer for AIDS Therapy

Retroviral-Mediated Gene Transfer into Bone Marrow Cells and T and B Lymphocytes

Research in the Field of Early Detection of Lung Cancer

Evaluation of cDNA Clones Related to Cancer Metastases

Transforming Growth Factor-α/Pseudomonas Exotoxin Hybrid Proteins

Cytokines for Enhancing Drug Delivery and Pharmacologic Action

Clinical Development of Taxol

Interleukin-2/Pseudomonas Exotoxin

Development of Methods for PCR Amplification of DNA Sequences Directly from Clinical Specimens

Human Cytochrome p-450 cDNA Expression and Mutagenesis

Clinical Development of PALA

Testing of Antigens for Improvement of IL-2 Therapy

Immunoglobulin and Immunotoxin Therapeutics for HIV Infections and AIDS

The Function of Two Novel, Inducible Proteins Secreted by Activated T Cells

Novel CD4 Targeted Anti-HIV Agents

Construction of a Multiwell Cell Settling Chamber

AIDS Vaccine

Single Chain Bispecific Antibody

Hair Follicle Cell Biology, Biochemistry and Molecular Biology

Human Papillomavirus (HPV) Infection and Cervical Dysplasia

Development of Nontoxic Aqueous Chemiluminescent Systems for Use in Photodynamic Therapy

RECENT/CURRENT NCI CRADA COLLABORATORS

Abbott Laboratories/Johns Hopkins University

American Cyanamid Company

Amgen Incorporated

Bristol-Myers Squibb

Cetus Corporation

Creative BioMolecules

Dow Chemical

Genetic Therapy, Inc.

Gentest Corporation

Hoffmann-LaRoche

Immuno-U.S.

Integrated Genetics

Lofstrand Labs

Merck

Molecular Oncology

Molecular Vaccines

Neuro Probe, Inc.

Sandoz Forschungsinstitut

U.S. Bioscience

Upjohn Company

Virogenetics

N92-22429

TECHNOLOGIES FOR THE MARKETPLACE FROM THE CENTERS FOR DISEASE CONTROL

Frances L. Reid-Sanden
Science Program Coordinator
Technology Transfer Office
Centers for Disease Control
1600 Clifton Rd., N.E.
Mailstop A20
Atlanta, Georgia 30333

R. Eric Greene
Technology Transfer Coordinator
Technology Transfer Office
Centers for Disease Control
1600 Clifton Rd., N.E.
Mailstop A20
Atlanta, Georgia 30333

Dolores M. Malvitz, DrPH
Program Analyst
Technology Transfer Office
Centers for Disease Control
1600 Clifton Rd., N.E.
Mailstop A20
Atlanta, Georgia 30333

ABSTRACT

The Centers for Disease Control, a Public Health Service agency, is responsible for the prevention and control of disease and injury. Programs range from surveillance and prevention of chronic and infectious diseases to occupational health and injury control. These programs have produced technologies in a variety of fields, including vaccine development, new methods of disease diagnosis, and new tools to ensure a safer work environment.

Development of a vaccine against hepatitis A, a common viral illness in day-care centers, is now possible due to techniques that produce large quantities of the viral agent. A recombinant rabies vaccine may assist in eradicating wildlife rabies. This vaccine is generated by inserting the rabies gene that elicits protective antibodies into a mammalian virus that does not produce disease in human beings. Similar technology has also produced the protein critical to the laboratory diagnosis of rabies, eliminating the current need for using infectious virus. Legionella species-specific monoclonal antibodies for the detection of legionellae in environmental samples and clinical specimens have been developed. A rapid method to diagnose human cysticercosis, a disease caused by consuming contaminated pork, is also available for commercial application. Finally, concern over worker safety has stimulated development of devices such as one that controls waste anesthetic gases in veterinary surgical units.

Consistent with the objectives of the Federal Technology Transfer Act, CDC is committed to transfer these and other technological innovations from the laboratory to the marketplace, so new products can be available to augment the control and prevention of disease and injury.

INTRODUCTION

The C enters for Disease Control (CDC) is one of seven Public Health Service agencies. Established in 1946, the Communicable Disease Center, as it was known then, grew out of the Office of Malaria Control in War Areas (MCWA) headquartered in Atlanta, Georgia. From its early work with state and local health officials to fight infectious diseases, CDC has grown into five centers/institutes. CDC has become a leader in:

1. prevention of disease, disability, and premature death caused by infectious and chronic diseases;

2. injury or disease associated with environmental, home, and workplace hazards; and

3. controllable risk factors such as poor nutrition, smoking, lack of exercise, high blood pressure, stress, and drug misuse.

As you know, in 1986, Congress passed the Federal Technology Transfer Act of 1986 (FTTA-86) to improve the link between the Federal laboratories' technology base and U.S. businesses. This law and related legislation authorized Federal laboratories to patent and exclusively license inventions to, and collaborate with, businesses on research and development. Until 1986, the principal methods used by CDC to transfer technology outside the Government were training, education, and information dissemination; CDC scientists presented information about technological developments in papers and at professional meetings and trained other scientists in new technologies.

In 1988 CDC established the Technology Transfer Office (TTO). TTO staff are responsible for the patenting and licensing of CDC inventions and, with CDC's Office of General Counsel (OGC), negotiating the language and licenses of the Center's Cooperative Research and Development Agreements (CRADAs), Biological Materials Licensing Agreements (BMLAs), and Materials Transfer Agreements (MTAs). Additionally, the TTO staff develop technology transfer training material, and conduct and sponsor seminars to educate and update CDC's scientific community on technology transfer issues. We also actively participate in the local biomedical/biotechnical community through our membership in the Clifton Corridor Council (CCC). The CCC, a Georgia-based non-profit organization, was founded in 1989 to promote Georgia's biomedical research institutions and to attract biotech companies to our state. Some of CCC's 34 institutional members are the American Cancer Society, Emory University School of Medicine, Georgia Institute of Technology, Morehouse School of Medicine, and the Medical College of Georgia. In addition, the CCC has more than 100 corporate and numerous individual members.

CDC's TTO staff have developed a technology licensing document highlighting our patented and patent-pending licensable technologies. This document is updated frequently and includes a statement on CDC's CRADA policy, as well as a model BMLA. Additionally, the CCC has put together a licensable technologies document, featuring available technologies from its member institutions. Both documents are available at our exhibit booth.

The Office of Technology Transfer (OTT), National Institutes of Health (NIH), coordinates the technology transfer activity for the PHS agencies, including NIH; the Alcohol, Drug, and Mental Health Administration (ADAMHA); the Food and Drug Administration (FDA); and CDC. As a part of their PHS-wide effort to offer "one-stop-shopping," OTT packages related groups of patented and patent-pending inventions from PHS laboratories, thus allowing licensing of complementary or interrelated technologies. OTT's profiles database is a mechanism whereby companies desiring to be contacted about new PHS inventions as they become available can share information about their corporate interests. Additionally, PHS-OTTO (Office of Technology Transfer On-Line) is a computer database providing modem access to data that can be copied into the user's own computer files. This electronic bulletin board includes descriptions of patents and patent applications available for licensing and brief descriptions of existing CRADAs. Names and telephone numbers of PHS scientists interested in CRADAs and their individual areas of expertise are highlighted. Also included are directories of PHS resource people, forum and conference announcements, model technology transfer agreements, and technology transfer guidelines. Standardization of documents used by PHS agencies has further contributed to the one-stop-shopping concept. By easing the transfer of technology into the marketplace, OTT's efforts benefit not only public health, but the U.S. economy as well.

From implementation of FTTA-86 through September 1991, CDC has processed more than 100 employee invention disclosures resulting in 44 filed patent applications and 3 issued patents. CDC currently has 27 active CRADAs and has negotiated 24 BMLAs. Combined income from these activities has been significant.

A sample of the technologies developed by CDC and available for licensing follow; they span the realm of infectious disease diagnostics and vaccine candidates to worker health and safety devices. Many other technologies are available or are in development; these eight examples have been selected to show the range and variety available from CDC.

1. A number of promising technologies associated with hepatitis A (HAV) vaccine development and diagnosis have been generated at CDC. Among other qualities, efficient yields of large quantities of hepatitis A virus are needed for the commercial development of vaccines and diagnostic tests for the disease. A vaccine candidate HAV strain can now be collected in appropriately large volumes.

2. An hepatitis A virus found in cynomolgus monkeys appears to be nonpathogenic for humans. Studies with this strain indicate that low-level infections characterized by minimal virus shedding may occur. Thus, this virus could be used for an "infection permissive" vaccine, resulting in the induction of lifelong immunity, no clinical illness, and very little shedding of virus into the environment.

3. A new, rapid, and sensitive diagnostic test to detect hepatitis A is now available. The test uses specific primers and the polymerase chain reaction (PCR) to detect hepatitis A in serum, food, or environmental samples.

4. An oral rabies recombinant vaccine has been developed using raccoon poxvirus as the expression vector. The vaccine confers protection against rabies to a number of wild and domestic animals. Because humans are not a natural host for raccoon poxvirus, other veterinary vaccines have been produced using this expression vector system.

5. Recombinant rabies N protein produced by the baculovirus expression vector system offers a safe substitute for infectious rabies virus now used in the direct immunofluorescent test to diagnose rabies. The protein can be used as an immunogen to produce antisera and as adsorbent material for a specificity control reagent. It may also have value in the development of rabies vaccines.

6. A set of three monoclonal antibodies has been developed that, alone or in combination, react specifically with a protein found in all *Legionella*. These antibodies do not react with bacteria from other genera and can be used as reagents to detect *Legionella* in environmental samples and clinical specimens.

7. Antibodies to *Taenia solium* directed against one of seven diagnostic larval antigens can be detected in serum or cerebrospinal fluid by an immunoblot assay. This test is useful in the diagnosis of active cases of neurological disease caused by the organism.

8. An example of a device available for licensing is one that enables removal of 95% of the waste anesthetic gases associated with small veterinary surgical units. These harmful substances are

removed by a relatively compact cartridge containing activated charcoal connected to a conventional anesthetic administration system. Gases of vaporized anesthetic substances are selectively directed through the activated charcoal without the need for motors, blowers, or other powered devices.

SUMMARY

CDC has made substantial, rapid progress in implementing provisions of the technology transfer legislation. CDC's achievement in this regard is significant, both because of what has been accomplished, and because of the opportunities offered by future activity. Three of CDC's five centers/institutes are active participants in technology transfer. Continuing education of CDC investigators in the technology transfer process, combined with their desire to control and prevent disease and injury offer fertile ground for additional collaborative research and the development of patentable technologies. NIH experts in the areas of patent and license execution have made CDC's task easier with a staff of patent and licensing portfolio managers who coordinate like technologies from participating PHS laboratories for licensing purposes. Electronic bulletin boards, computerized databases, licensing documents, and conferences, both at CDC and OTT, are additional tools being used to expedite the transfer of technology.

The technologies presented today represent a growing commitment by CDC to put into practical use those inventions developed by its scientific and engineering staff to help ensure a better quality of life through more accessible health care products and processes.

N92-22430

ENHANCEMENT OF BIOLOGICAL CONTROL AGENTS FOR USE AGAINST FOREST INSECT PESTS AND DISEASES THROUGH BIOTECHNOLOGY

James M. Slavicek
Forest Service, USDA
Northeastern Forest Experiment Station
Forestry Sciences Laboratory
359 Main Road
Delaware, Ohio 43015

ABSTRACT

Research and development efforts in our research group are focused on the generation of more efficacious biological control agents through the techniques of biotechnology for use against forest insect pests and diseases. Effective biological controls for the gypsy moth and for tree fungal wilt pathogens are under development. The successful use of Gypchek, a formulation of the Lymantria dispar nuclear polyhedrosis virus (LdNPV), in gypsy moth control programs has generated considerable interest in that agent. As a consequence of its specificity, LdNPV has negligible adverse ecological impacts compared to most gypsy moth control agents. However, LdNPV is not competitive with other control agents in terms of cost and efficacy. We are investigating several parameters of LdNPV replication and polyhedra production in order to enhance viral potency and efficacy thus mitigating the current disadvantages of LdNPV for gypsy moth control, and have identified LdNPV variants that will facilitate these efforts. Tree endophytic bacteria that synthesize antifungal compounds have been identified and an antibiotic compound from one of these bacteria has been characterized. The feasibility of developing tree endophytes as biological control agents for tree vascular fungal pathogens is being investigated.

INTRODUCTION

Chemical pesticides and fungicides are the preferred control agents for forest insect pests and fungal diseases. In excess of 350 billion pounds of these agents are used annually in the United States to control pests and diseases in forestry, agriculture, and in residential areas. Rachel Carson's Silent Spring, focused widespread attention on the environmental hazards linked to pesticide use. Broad spectrum insecticides and fungicides have adverse impacts not only on their target organisms but also on beneficial insects and fungi, and consequently on the entire ecosystem. In addition, chemical residues may cause health problems among the human population. Interest in biological insect and fungal control agents is growing as a consequence of these concerns regarding chemical pesticide use. A number of bacteria synthesize antifungal compounds, and over 1500 microorganisms or microbial products have been identified that are insecticidal. Generally, natural control agents have little adverse ecological impacts due to their specificity for the target host. Long term environmental hazards and health concerns are not a factor with biological control agents since chemical residues are not present. Unfortunately, biological control agents suffer from several disadvantages in comparison to chemical pesticides, including cost of production, efficacy, and stability. The techniques of biotechnology offer a means of mitigating some of the disadvantages inherent in biological control agents. Efforts in our research work unit focus on the enhancement of an insect virus and a bacteria for control of a defoliating moth and fungal tree pathogens.

GYPSY MOTH CONTROL THROUGH AN INSECT VIRUS

The Lymantria dispar nuclear polyhedrosis virus (LdNPV), which is pathogenic to Lymantria dispar, the gypsy moth, was selected as a model system for the enhancement of a biological control agent for a forest insect pest. The gypsy moth was imported from Europe into North America near Boston in 1869. Since then the area of gypsy moth infestation has increased to include almost the entire New England area, New York, Delaware, Maryland, New Jersey, Pennsylvania, Virginia, West Virginia, Ohio, and Michigan (1). Several chemical insecticides have been used for gypsy moth control, including DDT which was one of the most effective.

Understanding of the environmental impacts of DDT and other chemical insecticides caused a shift to the use of primarily Dimilin and Bacillus thuringiensis (Bt) for current gypsy moth control efforts. LdNPV has the significant advantage over other control agents of specificity for the gypsy moth. Consequently, LdNPV is the agent of choice for use in environmentally sensitive areas. LdNPV is not used extensively for gypsy moth control primarily as a consequence of high production costs and low efficacy. Efforts are being made to mitigate these problems inherent in LdNPV by generating viral strains that are competitive in terms of cost and efficacy with other gypsy moth control agents.

LdNPV replication and production for field use.

Two forms of LdNPV are produced during viral replication, occluded and nonoccluded virus (2). Early after infection nonoccluded virus is produced that leaves the cell and is responsible for secondary infections within the hemocoel of the host insect. A different form of virus is produced late in infection that is occluded into a protein matrix composed primarily of a viral encoded protein termed polyhedrin. These structures, approximately 1 to 3 micrometers in diameter and polyhedral in shape, are termed polyhedral occlusion bodies or polyhedra. Nucleocapsids within polyhedra are protected from most environmental conditions with the exception of ultraviolet light. Virus is applied through sprayers in the field in the form of polyhedra, which are ingested by gypsy moth larvae feeding on tree foliage. Once within the alkaline environment of the insect midgut, polyhedra dissolve, releasing nucleocapsids which infect the insect midgut cells thereby initiating the infection process.

LdNPV is currently produced in gypsy moth larvae at a cost of approximately $30.00 for enough polyhedra to treat an acre of forest. Production is the most expensive component inherent in the use of LdNPV. Limitations exist in increasing the efficiency of production of polyhedra in larvae. Efforts to produce polyhedra in cell culture systems are being made to generate a more economical production methodology. In comparison, production costs of Dimilin and Bt are approximately $3.00 per acre equivalent. LdNPV can be made cost competitive by either increasing viral potency, thereby allowing treatment with fewer polyhedra per acre, or decreasing production costs. We have initiated investigations to gain an understanding of polyhedra production and the molecular basis for viral potency with the goal of enhancing polyhedra production in cell culture and viral potency.

Enhancement of LdNPV polyhedra production and potency.

A LdNPV variant has been identified that differs from wild type virus in several characteristics, including polyhedra production and potency (3). This variant, isolate 5-6, was isolated after approximately twenty passages in cell culture (in Lymantria dispar 652Y cells, 4). In comparison to wild type LdNPV (isolate A21), isolate 5-6 produces fewer polyhedra both in cell culture and in vivo (Table 1).

Table 1. Polyhedra production in fourth instar gypsy moth larvae and in 652Y cells.

Isolate	Polyhedra/larvae	Polyhedra/652Y cell
5-6	8.6×10^7	4.4 ± 0.9
A21	2.1×10^9	51.0 ± 6.0

The relative number of virions present within polyhedra synthesized by these isolates was investigated. Polyhedra were produced in gypsy moth larvae and prepared for examination by electron microscopy. Polyhedra were sectioned, and the number of virions present in cross sections was quantified by counting and expressed as the number of virions per square micrometer of polyhedra cross section surface area (Table 2). Polyhedra produced by isolate 5-6 were found to be almost devoid of occluded nucleocapsids. The relative potency of LdNPV isolates 5-6 and A21 were examined through bioassay in second instar gypsy moth larvae (Table 3). Isolate 5-6 was found to exhibit very low potency in comparison to isolate A21.

Table 2. Number of virions present within cross sections of polyhedra produced by isolates 5-6 and A21 in fourth instar gypsy moth larvae.

Isolate	# of Polyhedra Cross Sections Examined	Average Diameter of Polyhedra Cross Sections	Total Number of Virions Counted	Number of Virions Per Square Micrometer
5-6	59	1.4 um	14	0.13
A21	25	2.0 um	941	12.2 ± 5.5

Production of few polyhedra, few virions occluded within polyhedra, and extremely low potency are traits exhibited by few polyhedra (FP) viral variants. FP variants have been identified in a number of baculovirus species (4), and have been found to arise during in vitro viral replication as a consequence of insertion of DNA sequences into preferential locations in the viral genome. FP mutations are generated with a high frequency during passage in cell culture, which places limitations on production of polyhedra in cell culture systems. Current investigations on isolate 5-6 are directed to an understanding of the molecular basis that gives rise to the FP phenotype. Once this is understood, a means of increasing the number of virions occluded into polyhedra may be devised which could enhance potency. In addition, if preferential DNA insertion sites exist that give rise to the FP phenotype, site directed mutagenesis could be used to create a viral strain lacking these sites which would preclude the genesis of the FP mutant. This viral strain would enhance cost effective polyhedra production in cell culture systems.

Table 3. Lethal Concentrations of A2-1, 5-6, and LDP 226 for second instar gypsy moth larvae.

Isolate	LC_{50}(95%FL)[a]	LC_{90}(95%FL)	Slope[b]	PR[c]
A2-1	9.90(7.29-13.47)	29.67(20.80-49.6)	2.68 ± 0.35	0.94
5-6	>7400	-	-	<0.0013
LDP 226	9.31(6.46-13.50)	48.61(30.37-97.42)	1.79 ± 0.23	-

a Polyhedra/ml diet X 10^3, FL= fiducial limits.
b ± SEM.
c Potency ratio = LC_{50} LDP 226/LC_{50} isolate.

Enhancement of LdNPV polyhedra production in cell culture.

Another characteristic of in vitro viral replication under study is the decrease in polyhedra production that occurs with viral passage. Cells in culture infected with nonoccluded virus isolated from hemolymph of virally infected gypsy moth larvae were found to produce the greatest number of polyhedra per infected cell (Table 4). After several passages in cell culture, the number of polyhedra produced per cell dropped approximately 3 fold. This observation has also been noted during the passage of other baculoviruses in cell culture, and appears not to be related to FP viral mutations. The basis for the observed decrease in polyhedra production is under investigation. A recently discovered LdNPV variant, MPV, may aid in the investigation of this characteristic. This isolate was obtained during passage of isolate A21 in cell culture, and was found to maintain a high level of polyhedra production after several passages in cell culture (Table 4). In addition, infection of cells with isolate MPV generates a greater percentage of cells containing polyhedra in comparison to isolate A21. Restriction endonuclease analysis of the genome of isolate MPV is being performed to determine if this isolate was generated by mutation of isolate A21. The economics of polyhedra production in cell culture would be enhanced if a means of maintaining a high level of polyhedra production over repeated passages could be devised.

Table 4. Polyhedra production in 652Y cells.

Isolate	1st Passage Production per Cell	Polyhedra Production Per Cell After More than 5 Passages	Percentage of Cells Containing Polyhedra
A21	51.0		38.0
A21		14.7	34.3
MPV		87.9	90.0

Identification of LdNPV genotypic variants exhibiting a range of potencies.

The identification and analysis of viral isolates exhibiting a range of biological activity may lead to a means of enhancing potency through a natural viral characteristic. Several LdNPV isolates were obtained from Gypchek, an LdNPV product currently used for control of the gypsy moth. Viral lines were generated through infection of gypsy moth larvae with amounts of polyhedra that result in from 5 to 10% larval mortality (6). At this mortality level, there is a high probability of generating an infection by a single virion or virion bundle which would generate a population of virus with one or a limited number of genotypes. The degree of genotypic heterogeneity in the LdNPV isolates was assessed by analysis of restriction endonuclease fragment polymorphisms after genomic restriction endonuclease digestion with the enzyme Bgl II. Gypchek was found to be composed of a number of LdNPV genotypes as evidenced by the identification of twenty-two genotypic variants. These variants were grouped into classes on the basis of similarities in the number and length of genomic fragments generated by digestion with the restriction enzyme Bgl II (Table 5).

The biological activity of some of the LdNPV isolates described above were determined through bioassay in gypsy moth larvae. The isolates were found to exhibit a range of potencies in relation to Gypchek, which was used as a standard (Table 6). Several of these isolates were selected for further studies designed to elucidate the basis for observed potency differences.

Table 5. Grouping of Variants According to Similarities in Genomic Bgl II restriction Patterns.

Class	Isolates	# Of Fragments	# Of Matching Fragments	Genomic Size
I	151	20	17	162.05
	A21			162.1
II	201	18-20	13	164.6
	203			164.65
	201-1			164.2
	203-1			163.5
	203-2			163.1
	203-8			163.15
	203-10			164.55
III	111	20	17	162.0
	163-3			161.8
	141-2			162.0
IV	123	19-20		160.3
V	121	19	16	161.6
	122			161.5
	141			161.7
	163			161.8
	B21			161.7
	B21-2			161.7
	131			161.7
VI	121-1	19		164.5
VII	162	20		163.6

Table 6. Relative potencies[1] of viral isolates.

ISOLATE	LC_{50}(LIMITS)[2]	LC_{90}(LIMITS)[2]	SLOPE (SEM)	PR_{50}[3]	PR_{90}[4]
203	1.5 (1.0- 2.2)	6.9 (4.4- 14.1)	1.9 (0.3)	4.4	4.6
A21	9.9 (7.3-13.5)	29.7 (20.8- 49.6)	2.7 (0.4)	0.9	2.0
201	5.9 (3.9- 8.8)	43.4 (25.7- 95.8)	1.5 (0.2)	1.8	1.5
111	8.8 (6.3-12.3)	35.7 (23.5- 67.0)	2.1 (0.3)	1.1	1.5
163	3.7 (2.2- 6.0)	61.3 (32.3- 155.4)	1.1 (0.1)	3.0	1.1
121	12.4 (8.9-17.3)	48.5 (32.2- 89.9)	2.2 (0.3)	0.8	1.1
226[5]	8.9 (6.0-13.3)	48.6 (33.3- 103.8)	1.8 (0.2)	1.0	1.0
151	7.7 (5.1-11.4)	52.8 (32.0- 108.2)	1.5 (0.2)	1.2	0.9
123	15.6(10.6-23.1)	102.3 (61.7- 212.7)	1.6 (0.2)	0.6	0.5
141	33.8(23.0-50.0)	216.1 (130.9- 444.8)	1.6 (0.2)	0.3	0.2
B21	21.8 (9.9-43.4)	277.1 (117.6-1132.4)	1.2 (0.1)	0.3	0.1
131	35.3(17.9-69.9)	648.9 (270.6-2606.3)	1.0 (0.1)	0.3	0.1

1. IN VIVO BIOASSAY IN 2ND INSTAR GYPSY MOTH LARVAE

2. POLYHEDRAL INCLUSION BODIES x 10^3 / ML DIET

3. RATIO OF LC_{50} STANDARD TO LC_{50} ISOLATE

4. RATIO OF LC_{90} STANDARD TO LC_{90} ISOLATE

5. STANDARD GYPCHEK PREPARATION

Enhancement of LdNPV efficacy.

Another disadvantage of LdNPV in comparison to chemical pesticides is that the virus requires from approximately 10 to 14 days to kill its host. During this time period, the virally infected larvae continue to feed, causing defoliation. Alterations to the virus that would either enhance viral killing speed or cause feeding cessation would provide greater protection to tree foliage. The efficacy of the Autographa californica nuclear polyhedrosis has been enhanced through the insertion of neurotoxin genes from the straw itch mite Pyemotes tritici (7) and the scorpion Androctonus australis (8) into the viral genome. In both of these cases, insects infected with the recombinant viruses died in less time compared to insects infected with wild type virus. The neurotoxin genes from P. *tritici* and A. *australis* will be inserted into the LdNPV genome and the efficacy of the recombinant viruses assessed through bioassay.

Through the approaches described in the previous sections, it is anticipated that LdNPV strains will be developed with enhanced control properties that mitigate current disadvantages of LdNPV for gypsy moth control, providing an ecologically sound control agent with commercialization feasibility.

CONTROL OF FUNGAL VASCULAR WILT TREE PATHOGENS

The fungal vascular wilt pathogens *Ophiostoma ulmi* and *Cryphonectria parasitica* have decimated American elm and chestnut trees in North America. Chemical fungicides have proven effective in control of these fungal pathogens when administered through injection. This labor intensive and expensive control strategy is acceptable on a limited basis, but is not feasible for treatment in a forest setting. Treatment of tree fungal pathogens is especially difficult due to the need for delivering the control agent to the vascular system of the tree, thus precluding aerial application methods.

As an alternative to chemical fungicides, the feasibility of using host tree bacterial endophytes is being assessed. Strains of *Bacillus* and *Pseudomonas* bacteria that synthesis antifungal compounds have been isolated from tree vascular systems. The structure and activity of an antibiotic of the iturin group has been characterized that is produced by B. *subtilis* (9,10). This antibiotic is effective against O. *ulmi* and C. *parasitica* in *in vitro* assays. If the bacteria is to provide effective fungal control either the concentration of B. *subtilis* resident in the xylem of the tree and/or the amount of the antibiotic produced by the bacteria need to be increased. Whether this is possible remains to be determined. Recent studies on control of fungal wilt pathogens by *Pseudomonad* species suggests that an elicitation of natural tree defense mechanisms was the primary determinant in overcoming fungal infection (11, R. Scheffer, personal communication). These results highlight the complex fungal-host tree interactions that occur upon infections, and suggest the need for obtaining a better understanding of these interactions. Additional research may lead to the development of efficacious and cost effective controls for tree vascular wilt pathogens that can be used in a forest setting.

REFERENCES

1) McFadden, M.W. and McManus, M.E. (1989) An insect out of control? The potential for spread and establishment of the gypsy moth in new forest areas in the United States. In Forest Insect Guilds: Patterns of Interaction with Host Trees, Y.N. Baranchikov, W.J. Mattson, F.P. Hain and T.L. Payne (eds.) pp. 172-186. US Department of Agriculture, For. Serv. Gen. Tech. Rep. NE-153, Radnor, PA. 1989.

2) Granados, R.R. and Williams, K.A. (1986) In vivo infection and replication of baculoviruses. In The Biology of Baculoviruses, R.R. Granados and B.A. Federici (eds.) pp. 89-108. CRC Press, Boca Raton, Florida, 1986.

3) Slavicek, J.M., Podgwaite, J. and Lanner-Herrera, C. (In press, 1992) Properties of two *Lymantria dispar* nuclear polyhedrosis virus isolates obtained from the microbial pesticide gypchek. J. Invertebr. Pathol.

4) Goodwin, R.H., Tompkins, G.J. and McCawley, P. (1978) Gypsy moth cell lines divergent in viral susceptibility. In Vitro 14, 485-494.

5) Fraser, M.J. (1987) FP Mutation of nuclear polyhedrosis viruses: A novel system for study of transposon-mediated mutagenesis. In Biotechnology in Invertebrate Pathology and Cell Culture, K. Maramorosch (ed.) pp. 265-293. Academic Press, San Diego, 1987.

6) Slavicek, J.M. and Podgwaite, J. Isolation of Lymantria dispar nuclear polyhedrosis virus genotypic variants exhibiting a range of biological activity. In Preparation.

7) Tomalski, M.D. and Miller, L.K. (1991) Insect paralysis by baculovirus-mediated expression of a mite neurotoxin gene. Nature 352, 82-85.

8) Stewart, L.M.D., Hirst, M., Ferber, M.L., Merryweather, A.T., Cayley, P.J. and Possee, R.D. (1991) Construction of an improved baculovirus insecticide containing an insect-specific toxin gene. Nature 352, 85-88.

9) Eshita, S.M., Roberto, N.H., Workman, R.F., Beale, J.M. and Mamiya, B.M. Bacillomycin L, A new antibiotic of the iturin group. I. Isolation and structure determination. In Preparation.

10) Eshita, S.M., Roberto, N.H. and Workman, R.F. Bacillomycin L, A new antibiotic of the iturin group. II. Effect of the beta-amino acid on antifungal activity. In Preparation.

11) Scheffer, R.J. (1989) Pseudomonas for biological control of Dutch elm disease. III. Field trials at various locations in the Netherlands. Neth. J. Pl. Path. 95, 305-318.

N92-22431

Use of T7 RNA Polymerase to Direct Expression of Outer Surface Protein A (OspA) from the Lyme Disease Spirochete, *Borrelia burgdorferi*.

John J. Dunn and Barbara N. Lade
Biology Department
Brookhaven National Laboratory
Upton, NY 11973

ABSTRACT

The *ospA* gene from a North American strain of the Lyme disease Spirochete, *Borrelia burgdorferi*, has been cloned under the control of transcription and translation signals from bacteriophage T7. Full-length OspA protein, a 273 amino acid (31kD) lipoprotein, is expressed poorly in *Escherichia coli* and is associated with the insoluble membrane fraction. In contrast, a truncated form of OspA lacking the amino-terminal signal sequence which normally would direct localization of the protein to the outer membrane is expressed at very high levels ($\geq$ 100mg /liter) and is soluble. The truncated protein has been purified to homogeneity and is being tested to see if it will be useful as an immunogen in a vaccine against Lyme disease. Circular dichroism and fluorescence spectroscopy has been used to characterize the secondary structure and study conformational changes in the protein. Studies underway with other surface proteins from *B. burgdorferi* and a related spirochete, *B. hermsii*, which causes relapsing fever, leads us to conclude that a strategy similar to that used to express the truncated OspA can provide a facile method for producing variations of Borrelia lipoproteins which are highly expressed in *E. coli* and soluble without exposure to detergents.

INTRODUCTION

Lyme disease is caused by the spirochete *Borrelia burgdorferi* and this complex disorder is the most common vector-borne infection in the USA. Specific diagnostic tests and ultimately a vaccine for Lyme disease are of major importance. In this regard, the outer surface lipoproteins (OspA and OspB) are of interest because they are unique to *B. burgdorferi* and not shared with other Borrelia or spirochetes. These proteins are highly immunogenic in experimental animals, however, the antibody response to these antigens occurs late in the course of natural infection and only in a small minority of patients (<5%) [1-3]. While the exact functions of these proteins are unknown, Osp-enriched fractions have been utilized as immunodiagnostics for late disease and in the design of "selective Westerns". The genes for these two proteins (*ospA* & *ospB*) are encoded in tandem in a 2KB stretch of a 49 kB linear plasmid separated one from the other by only 12 base pairs and co-transcribed from the same promoter [4, 5]. The deduced amino acid sequences of OspA and OspB begin with N-terminal sequences believed to constitute signal sequences which normally direct localization of the protein products to the spirochetes outer membrane.

We have chosen OspA as a model system for studying expression of *B. burgdorferi* proteins, particularly surface proteins, in *E. coli*. These proteins most likely affect the antigenicity, immunological reactivity, host cell interactions of the spirochete, invasion of the host, and development of symptoms; information about these proteins is expected to be useful in the diagnosis, treatment and prevention of the disease. Our goal is to use recombinant DNA techniques to overexpress these proteins to standardize immunoassays and other diagnostic screening tests and to serve as potential antigens for a vaccine.

METHODS

T7/pET system

B. burgdorferi OspA protein was expressed using the T7/pET expression system developed at Brookhaven [6] in which the RNA polymerase of bacteriophage T7 is used to drive transcription of a cloned target sequence. Bacteriophage T7 RNA polymerase is a single-polypeptide enzyme that is highly selective for specific promoters that are rarely encountered in DNA unrelated to T7 DNA. Efficient termination signals are also rare, so that T7 RNA polymerase is able to make complete transcripts of almost any DNA that is placed under control of a T7 promoter. A very active enzyme, T7 RNA polymerase elongates chains about five times faster than does *Escherichia coli* RNA polymerase. These properties, together with the availability of the cloned gene have been exploited as the basis of expression systems in *E. coli* and other cell types, including mammalian and yeast cells.

Target genes are cloned in pET plasmids downstream from a strong T7 promoter and adjacent to efficient translational initiation signals from the phage. A host strain containing a chromosomal copy of the gene for T7 RNA polymerase under control of the IPTG-inducible *lac*UV5 promoter is used for protein production. The T7 polymerase is so selective and active that, after IPTG induction, almost all of the cell's resources are converted to target gene expression; after a few hours, the desired protein product can constitute more than 50% of the total cell protein. The host strain we use, BL21(DE3)/pLysS, also contains a plasmid which specifies low levels of T7 lysozyme, a natural inhibitor of T7 RNA polymerase. In uninduced cells, lysozyme reduces the basal activity of the T7 RNA polymerase and increases the range of target genes that can be stably maintained in the expression host. We used a pET vector, pET9 that has a *kan* gene, conferring resistance to the antibiotic kanamycin, as the selectable marker instead of the more commonly used pET vectors with a gene conferring resistance to ampicillin, *bla*. Ampicillin is not used during growth, so ampicilloyl-target protein conjugates cannot be formed. Such conjugates would complicate projected immunological studies [7]. The host *E. coli* strain, BL21, lacks the *ompT* outer membrane protease that can degrade proteins during purification [8], and it is deficient in the *lon* protease that can degrade recombinant proteins as they are being synthesized [9].

Polymerase Chain Reaction

Polymerase chain reaction (PCR) technology [10] has made cloning of genes into the pET vectors efficient and straightforward. Not only does PCR permit targeted amplification of specific DNA sequences starting with very small quantities of template, but also it provides an efficient means to fuse signals, such as new, unique restriction sites, to the 5' and 3' ends of a target sequence. Synthetic oligonucleotide primers are designed to be complementary to a limited portion, typically about 15 nucleotides, at each end of the target sequence and to include the desired restriction enzyme recognition sequences near their 5'- ends. During amplification the DNA sequences of the primers are incorporated into the product, thereby, allowing new sequence information in the 5' ends of the primers to be fused to the target sequence. After cutting with the appropriate restriction endonuclease, the resulting amplified DNA is ready for insertion directly into the pET expression plasmid. The natural ATG initiation codon in the vector is part of an *Nde*I site (CA|TATG), so coding sequences are easily inserted directly at the initiation codon by including an appropriately positioned *Nde*I site in the forward PCR primer. Other vectors in the pET set have an *Nco*I site (C|CATGG) at their initiation codon in case the target gene has an *Nde*I site(s) in its sequence. PCR-mediated cloning also can be used to split a gene into regions that code for individual domains and antigenic epitopes and these segments can be expressed individually to eliminate cross-reacting antigenic domains. It is important to note that PCR amplification can be used to eliminate endogenous *B. burgdorferi* promoter signals in front of a protein coding sequence. In all likelihood these prokaryotic promoters would be recognized by the host polymerase and the plasmids containing them would be unstable due to significant levels of basal transcription of the foreign gene by *E. coli* RNA polymerase. Too high a level of unregulated basal expression probably explains why Isaacs and Radolf [11] were unable to clone the endoflagellar sheath protein gene from a related spirochete, *Treponema pallidum*, in *E. coli* if they included the native promoter in front of the coding sequence.

RESULTS

We have used two sequence specific sets of oligonucleotide primers to amplify and clone related forms of the *ospA* gene under the control of transcription and translation signals from T7 [12]. One set allowed the entire ospA sequence to be cloned, while the other primed amplification of a truncated form of *ospA* lacking the first 17 codons specified by the wild-type structural gene, *i.e.* the residues believed to act as a signal sequence to direct association of OspA with the Borrelia membrane. On the basis of its similarity with *E. coli* signal sequences it seemed possible that the full-length OspA protein would be lipidated in *E. coli* and become membrane associated. Typically such proteins have poor solubility properties and consequently detergents are required to solubilize them.

The recombinant version of the full-length *ospA* gene was expressed at a low level, a few % of the total protein, even when the coding sequence was placed under control of the strong transcription and translation signals from bacteriophage T7 in the pET vector. The low level of expression presumably is due to the accumulated toxic effects of OspA protein localizing at the *E. coli* cell membrane during expression and, as expected, the full-length protein remains associated with the insoluble *E. coli* membrane fraction unless solubilized by detergents. In contrast, the truncated form of the OspA protein lacking the 17 amino acid long signal sequence is expressed at very high levels, greater than 50% of the total protein, and it is highly soluble. We have developed an efficient procedure for purifying large amounts of recombinant OspA to homogeneity; more than 100 mg of protein can be obtained from 1 liter of induced cells. The recombinant OspA is highly soluble in aqueous solution (≥ 50 mg/ml) and it elutes from a gel filtration column as a monomer under nondenaturing conditions. The molar extinction coefficient at 280 nm of purified OspA has been determined, simplifying quantitation in immunological testing.

The recombinant protein begins with two amino acids from the vector (Met Ala) fused in frame to Lysine18 of the OspA sequence. Amino-terminal sequencing of purified protein demonstrates that the first methionine residue is removed, therefore, the recombinant protein is referred to as OspA-257. Western blotting demonstrated that, although the OspA-257 protein is missing a lipidated N-terminus, it still reacts with antibodies present in the synovial fluid of a patient with Lyme-induced arthritis and is recognized by a variety of polyclonal and monoclonal antibodies raised against whole *B. burgdorferi* cells. Thus, it seems highly likely that the protein will be useful in immunochemical analysis for detection of Lyme disease.

We of course would like to know the three-dimensional structure of OspA since such information would assist in elucidation of antigenic sites and eventually may lead to development of a peptide-based vaccine. X-ray diffraction is still the only method for determining the exact three-dimensional structure of proteins in the size range of OspA and attempts to crystallize OspA are in progress. In lieu of a crystal structure, we are using other methods to characterize the structure of OspA such as circular dichroism (CD) spectroscopy. Unlike X-ray diffraction, CD can be applied to proteins in solution, under physiological conditions. Each type of secondary structure is characterized by a distinctive CD spectrum, and recent work by Johnson [13] showed that when CD spectra of proteins are extended below 200 nm to 178 nm, five types of secondary structure can be predicted with accuracy comparable to X-ray diffraction: α-helix, anti-parallel β-sheet, parallel β-sheet, β-turns (all types) and aperiodic structures (including random coil).

We used the CD spectrometer at the National Synchrotron Light Source at BNL to measure the CD spectra of OspA to wavelengths as short as 175 nm; conventional CD spectrometers measure spectra reliably in the wavelength region ≥ 200 nm. The spectra were analyzed with a computer program supplied by C.W. Johnson, Jr. This algorithm uses matrix techniques (singular value decomposition) and statistical procedures (variable selection) to fit the CD spectra of the protein to a linear combination of orthogonal CD spectra derived from a library of reference proteins, whose secondary structures are known from X-ray diffraction. The results [14] show that OspA contains mostly β-sheet (27% anti-parallel, 9% parallel) configurations, β-turn (21%), and random coil (34%), but little α-helix (11%). These values are quite different from those obtained using purely predictive methods based solely on the amino acid sequence of OspA, which predict a much higher percent of α-helix than is present (Table 1).

Description of Method	Type of structure %		
	α-helix	ß-all types	others
UV-CD (our work)	11±1	57±2	34±1
PCGene	24 (27)*	22 (21)	54 (52)
Chou-Fasman	45 (40)	28 (34)	27 (26)
Garnier-Robson	54 (53)	17 (18)	29 (29)

* (predicted values for full-length, unprocessed OspA)

TABLE 1.

We have also used steady-state and time-resolved fluorescence to investigate the static and dynamic aspects of OspA conformation. OspA contains a single tryptophan residue at position 200 (residue 216 in the full-length protein). Tryptophan can be selectively excited at wavelengths ≥ 295 nm and its fluorescence emission is particularly sensitive to the local environment. The fluorescence signal provides a probe for changes in protein conformation since changes in protein conformation are expected to cause changes in the environment of tryptophan. The emission peak (330 nm) of OspA and its response to various ionic and nonionic quenchers indicate that Trp200 is buried within the native protein in a relatively hydrophobic environment. Trp200 is completely exposed to the solvent, as expected for a random coil, when OspA is denatured by high temperature (≥ 80°C) or guanidine.

Both CD and fluorescence measurements reveal that the native conformation of OspA is highly stable; no significant changes are seen from pH 3-11, and the protein is stable at low and high salt concentrations. The protein returns to its native state, as reflected by CD and fluorescence measurements, after being held at 100°C for up to 10 min. Interestingly, the OspA-257 protein is exceptionally resistant to digestion by trypsin and human plasmin, even though it is rich in lysine residues (16 mol %) which are distributed rather uniformly throughout the protein, but it is extremely susceptible to proteinase K.

The same strategy used to clone *ospA* was used to clone recombinant versions of *ospB* directly into a pET vector with similar results, *i.e.* the full-length lipidated protein is poorly expressed while its truncated counterpart is overproduced[1]. The truncated OspB protein is currently being characterized by CD and fluorescence; it also has a single tryptophan.

In a related spirochete, *B. hermsii*, antigenic variation is the consequence of sequential expression of genes for a set of variable outer surface proteins known as Vmps [15]. Recently, it has been demonstrated that the Vmp genes have many features in common with the *B. burgdorferi* Osp gene family including N-terminal sequences that are typical of lipoproteins and evidence has been presented that the Vmp proteins are recognized and processed as lipoproteins in *E. coli* [16]. As might be expected, full-length Vmp's are expressed at relatively low levels in *E. coli*[2] presumably because the protein is toxic when it becomes anchored in the membrane. Vmp overproduction was possible if the region of the gene encoding the hydrophobic signal sequence was excised before the recombinant gene was placed under control of T7 signals in a pET plasmid[3]. Moveover, the truncated Vmp protein is soluble and it can be readily purified to homogeneity without the use of detergents.

[1] Duray, P., Lade, B.N. and Dunn, J.J. (unpuplished).
[2] Barbour, A.G. (personal communication).
[3] Bundoc, V.G., Barbour, A.G., Lade, B.N. and Dunn, J.J. (unpublished)

SUMMARY

The T7 based pET expression system is a powerful and versatile system for cloning and expression of recombinant proteins in *E. coli*. PCR technology complements the pET system since it permits precise, sequence specific amplification of coding sequences which, when used in conjunction with appropriately designed primers allows amplification of DNA tailored for direct insertion into a pET vector. The net result is a more highly efficient method to rationally engineer and express open reading frames. In certain instances, the target protein may contain sequences which limit overproduction. If the sequence is a specific domain of the target protein then it may be possible to remove it without destroying the protein per say. A good example of such a domain is the class of hydrophobic leader signal sequences found at the beginning of Borrelia lipoproteins (OspA, OspB and the Vmp's). Elevated expression of the full-length proteins is toxic to the overproducing strains and they stop making the protein after a short period of time. In contrast, these same proteins without the hydrophobic leader sequences continue to be synthesized at high rates for hours following induction and they accumulate to very high levels. The ability to produce these recombinant proteins in large amounts and in a highly purified state without using detergents raises the possibility that they may be useful in immunoassays and as possible immunogens in vaccines. Such studies are currently in progress.

REFERENCES

1. Craft, J.E., Filcher, D.K., Shimamoto, G.T., and Steere, A.C. (1986). Antigens of Borrelia burgdorferi recognized during Lyme disease: appearance of a new immunoglobulin M response and expansion of the immunoglobulin G response late in the illness. J. Clin. Invest. **78**, 934-939.

2. Magnarelli, L.A., and Anderson, J.F. (1988). Enzyme-linked immunosorbent assays for the detection of class-specific immunoglobulins to *Borrelia burgdorferi*. Am. J. Epidemiol. **127**, 813-815.

3. Luft, B.J., and Dattwyler, R.J. (1989). Lyme Borreliosis. In Current Clinical Topics in Infectious Disease. Volume 10. J. S. Remington, M. N. Swartz (Editors), Raven Press, New York, NY.

4. Howe, T.R., LaQuier, F.W., and Barbour, A.G. (1986). Organization of genes encoding the outermembrane proteins of Lyme disease agent *Borrelia burgdorferi* within a single transcriptional unit. Infect. Immun. **54**, 207-211.

5. Bergström, S., Bundoc, V.G., and Barbour, A.G. (1989). Molecular Analysis of linear plasmid-encoded major surface proteins, OspA and OspB, of the Lyme disease spirochete *Borrelia burgdorferi*. Mol. Microbiol. **3**, 479-486.

6. Studier, F.W., Rosenberg, A.H., Dunn, J.J., and Dubendorff, J.W. (1990). Use of T7 RNA polymerase to direct expression of cloned genes. Meth. Enzymol. **185**, 60-89.

7. Yvon, M., Anglade, P., and Wal, J-M. (1990). Identification of the binding sites of benzyl penicilloyl, the allergenic metabolite of penicillin, on the serum albumin molecule. FEBS **263**, 237-240.

8. Grodberg, J. and Dunn, J.J. (1988). *ompT* encodes the *Escherichia coli* outer membrane protease that cleaves T7 RNA polymerase during purification. J. Bacteriol. **170**, 1245-1253.

9. Goff, S.A. and Goldberg, A.L. (1985). Production of abnormal proteins in E. coli stimulates transcription of *lon* and other heat shock genes. Cell **41**, 587-595.

10. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., and Erlich, H.A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **239**, 487-491.

11. Isaacs, R.D., and Radolf, J.D. (1990). Expression in *Escherichia coli* of the 37-Kilodalton endoflagellar sheath protein of *Treponema pallidum* by use of the polymerase chain reaction and a T7 expression system. Infect. Immun. **58**, 2025-2034.

12. Dunn, J.J., Lade, B.A., and Barbour, A.G. (1990). Outer Surface protein A (OspA) from the Lyme Disease Spirochete, *Borrelia burgdorferi*: High Level Expression and Purification of a Soluble Recombinant Form of OspA. Protein Expression and Purification **1**, 159-168.

13. Johnson, W.C., Jr. (1990). Proteins: structure, function and genetics. **7**, 205-214.

14. France, L. L., Kieleczawa, J., Dunn, J. J., Hind, G., and Sutherland, J. C. Structural Analysis of an Outer Surface Protein from the Lyme Disease Spirochete, *Borrelia burgdorferi*, using Circular Dichroism and Fluorescence Spectroscopy. Biochem. Biophys. Acta. (in press).

15. Barbour, A.G., and Stonner, H.G. (1984). Antigenic variation of *Borrelia hermsii*. UCLA Symp Mol Biol New Ser **20**, 123-135.

16. Burman, N., Bergström, S., Restrepo, and Barbour, A.G. (1990). The variable antigens Vmp7 and Vmp21 of the relapsing fever bacterium *Borrelia hermsii* are structurally analogous to the VSG proteins of the African trypanosome. Mol. Microbiol. **4**, 1715-1726.

COMMUNICATIONS

(Session A3/Room C4)

Tuesday December 3, 1991

- Commercial Applications of ACTS Mobil Terminal Millimeter-Wave Antennas
- Antennas for Mobile Satellite Communications
- MMIC Linear-Phase and Digital Modulators for Space Comunications Applications
- Phased-Array Antenna Beamforming Using an Optical Processor

N92-22432

COMMERCIAL APPLICATIONS OF THE ACTS MOBILE TERMINAL MILLIMETER-WAVE ANTENNAS

Arthur C. Densmore, Rick A. Crist, Vahraz Jamnejad, Ann N. Tulintseff
JPL, California Institute of Technology
Pasadena, CA 91109

ABSTRACT

NASA's Jet Propulsion Laboratory is currently developing the Advanced Communications Technology Satellite (ACTS) Mobile Terminal (AMT), which will provide voice, data and video communications to and from a vehicle (van, truck, or car) via NASA's geo-stationary ACTS satellite using the K- and K_a-band frequency bands. The AMT is already planned to demonstrate a variety of communications from within the mobile vehicular environment, and within this paper a summary of foreseen commercial application opportunities is given. A critical component of the AMT is its antenna system, which must establish and maintain the basic RF link with the satellite. Two versions of the antenna are under development, each incorporating different technologies and offering different commercial applications.

INTRODUCTION

The ACTS satellite is scheduled for launch in 1993, and the first demonstrated use of the satellite will be made by the AMT [1]. From a van and a car that will travel along roads and highways a demonstration of voice, fax, and compressed video communication will be made. The communication is two-way, between the vehicle in Los Angeles, California and a base station in Cleveland, Ohio via the ACTS satellite as shown in Figure 1. The AMT antenna which mounts to the vehicle roof with a diameter of 8 inches and height 3 inches must track the direction to the satellite as the vehicle moves about. The antenna is designed with a wide enough elevation coverage so that only azimuthal tracking is required, and a thin microprocessor-controlled, pancake-shaped motor drives the antenna azimuth angle. Two versions of the antenna are being developed and both will be the same size and use the same motor for tracking.

The requirements to which the antennas are being designed ensure that the overall AMT will provide a high performance measure. In addition to simply operating in a mobile environment for which shock and vibration are characteristic, the antenna must adhere to both RF and tracking requirements.

For the uplink transmit link from the mobile vehicle, the AMT must generate 22 dBW EIRP minimum power. The transmit power density must be large enough to accommodate the required communications rates but low enough that the transmit beam is not a hazard for people standing near the vehicle. For the downlink receive link to the vehicle the AMT must provide sufficient receive sensitivity to accommodate the downlink communications rates. The minimum requirement is -8 dB/K, and typically losses within the antenna must be minimize to make the sensitivity this high. The tracking requirements are that the AMT compensate for a maximum vehicle yaw (turn) rate of 45 deg per sec. When the communications link has not yet been established, and the AMT must acquire the satellite, the requirement is that acquisition be complete and the link established within 10 sec. While the vehicle is traveling and the antenna tracking the satellite, the required tracking accuracy is 0.8 deg rms. Since the ACTS satellite will be in geostationary orbit, the AMT tracking system must achieve essentially an inertially-stabilized antenna pointing system. As the host vehicle turns, the position of the antenna is adjusted to insure that the antenna continuously points towards the satellite. An embedded microprocessor controls the antenna's movements.

REFLECTOR ANTENNA

The reflector antenna is the first version of the AMT antenna to be developed. It represents a relatively low risk development for the short available schedule. The reflector requires that the AMT incorporate TWT (tube) technology, in order to meet the uplink transmit EIRP requirement. A single TWT is employed to generate the 1-3

 PRECEDING PAGE BLANK NOT FILMED

watts required to maintain the mobile satellite uplink. Although not normally required, the reflector antenna is capable of handling the full 15 watts that the TWT can generate at 30 GHz. A single low noise amplifier (LNA) unit with 2.5 dBNF ensures a sufficient receive system sensitivity at 20 GHz. Waveguide plumbing connects the reflector antenna to the TWT and LNA with minimum loss, so that the overall RF performance is not significantly degraded.

The reflector shown in Figure 2 consists of a waveguide feed horn assembly and a reflecting surface. The feed assembly consists of a diplexer, orthomode transducer (OMT), and a feed horn, which faces the reflecting surface. The diplexer has a coaxial port which connects to the RF transceiver of the AMT through a rotary joint, and two waveguide ports (REV & XMT) which connect to the OMT. The OMT combines the signals from the two diplexer ports into a single waveguide port with the proper waveguide modes at the two ACTS frequencies so that a single feed horn may be used to properly illuminate the reflector surface. The reflector surface is a section of a parabolic surface of revolution, with its boundary defined by the intersection with a cylindrical ellipsoid. The projected view of the reflecting surface appears nearly elliptical both from the view of the feed horn and the direction of propagation. The reflector is only a few wavelengths across both its dimensions, and so is electrically much smaller than the simple reflector antenna application, and its beam less collimated in the same regard.

COMMERCIAL APPLICATIONS OF THE REFLECTOR ANTENNA

Commercial applications which would make use of the reflector antenna are those which need to rely on technology which has already been extensively tested and proven. The reflector with its TWT & LNA represent a classical system which has for decades been used and proven in radar applications, although the AMT does not utilize these technologies in the same (radar) manner.

One such application would be remote news gathering vehicles which now depend on cellular systems for communication in urban centers; those systems do not operate in rural areas and could be supplemented by use the AMT system. This application would support video and audio from the most remote rural sites, world-wide. This service would not complete with existing cellular telephone systems, but supplement them instead. Within metropolitan areas more use would be made of the cellular system, while the direct satellite link could be exploited in rural areas.

Another such application might be emergency communication vehicles, which could respond immediately to disaster sites and provide needed communication via satellites independent of local power outages.

PLANAR ARRAY ANTENNA WITH MMICs

The planar array antenna will be developed after the reflector antenna, although its design has already begun. This antenna offers MMIC technology which is more advanced, yet less proven, than that of the reflector antenna.

The planar array antenna shown in Figure 3 utilizes solid state MMIC transmit (power) and receive (low-noise) amplifier circuits. These MMICs are integrated into the antenna on the array itself so that minimal transmission line length exists between the MMIC LNAs and the array antenna elements. The short lengths of transmission line keep the losses very low and the antenna efficiency and sensitivity high. The array is organized into 12 columns (groups) of antenna elements, and each column is directly connected to a receive or transmit amplifier. This arrangement provides for a direct spacial combination of the receive signal or transmit power. In this manner the transmit power amplifier in each column must only provide a fraction of the power that is radiated by the entire array, and this power level ($\approx$ 0.1 watt) can be supplied by MMIC circuits. Since the MMICs are integrated very near the antenna elements, the short interconnecting transmission lines present a smaller amount of loss, and the antenna as a whole is more power efficient than an antenna system which must distribute all transmit power from a single high-power source through longer transmission lines to each of the antenna elements.

COMMERCIAL APPLICATIONS OF THE PLANAR ARRAY WITH MMICs

Commercial applications of the planar MMIC antenna require a more costly initial investment that the reflector because of the need to achieve mass production of a presently more expensive technology. Once mass production is established and a market is available, it is expected that the commercial markets will eventually allow the overall costs to reduce even below that of the reflector antenna.

One exciting commercial application is in the new and growing field of personal communications. Today a wrist watch can be purchased that contains a beeper to notify the user of calls. The planar MMIC array is a first step toward the development of a miniaturized personal satellite communication handset. Such a handset would allow communication anywhere in the world.

Because of the modular design of the planar MMIC array antenna, its MMIC circuits may be readily enhanced to include electronic phase control, allowing phase-array electronic azimuthal steering, without having to modify the basic antenna design concept. Such a commercial application would be capable of tracking a high-speed vehicle such as a plane flying across the field of view at a relatively constant elevation angle.

CONCLUSION

The launch of the ACTS satellite will bring about new commercial opportunities, which the AMT antennas will begin to demonstrate, the AMT reflector antenna provides a system based on proven technologies, and the AMT planar MMIC array provides a "state-of-the-art" antenna system which points the direction to the future of personal, hand-held satellite communication systems.

ACKNOWLEDGEMENT

The research described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

REFERENCES

[1] K. Dessouky, et al., "The ACTS Mobile Terminal," JPL SATCOM Quarterly, No. 2, July 1991.

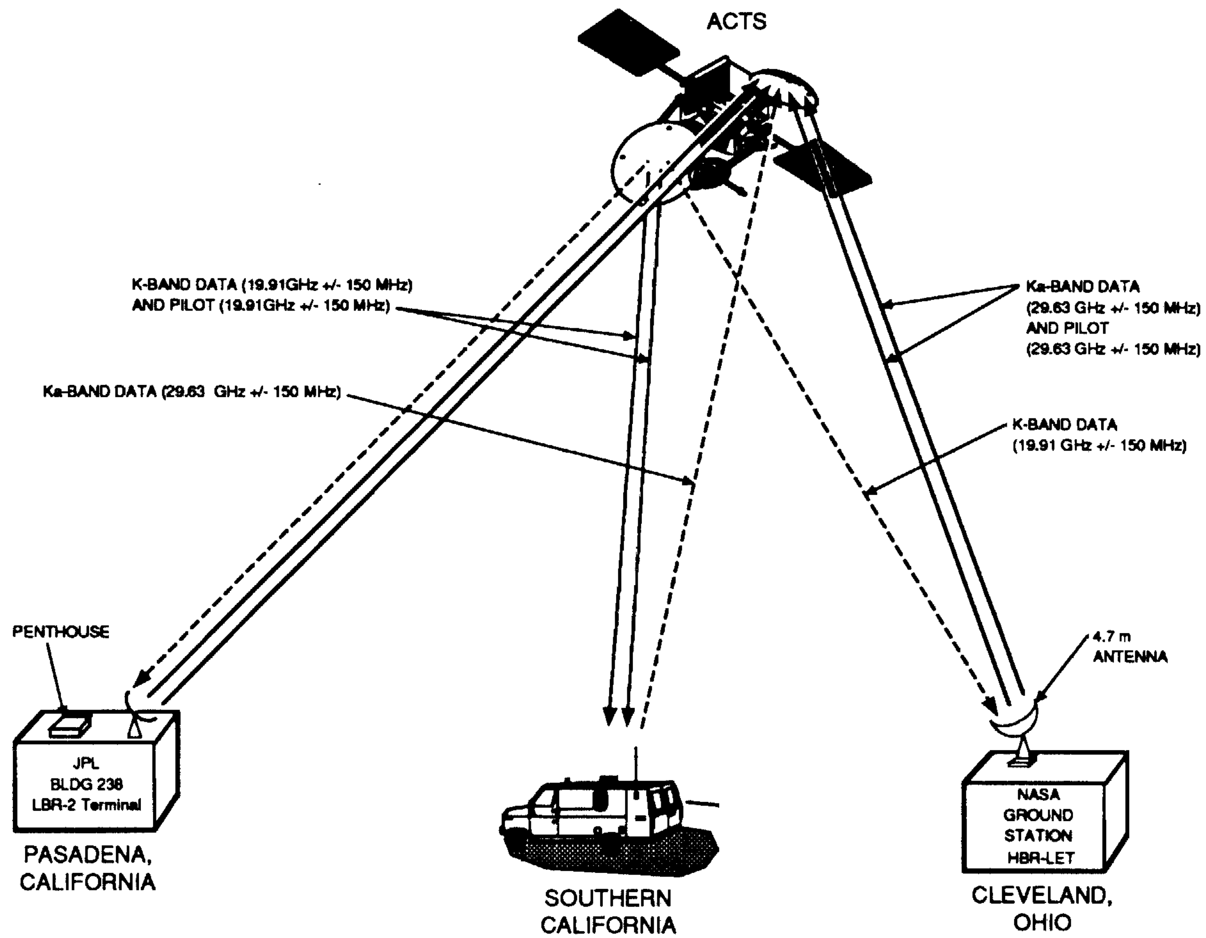


Figure 1. AMT experimental setup with millimeter-wave antennas mounted on top of mobile vehicle.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

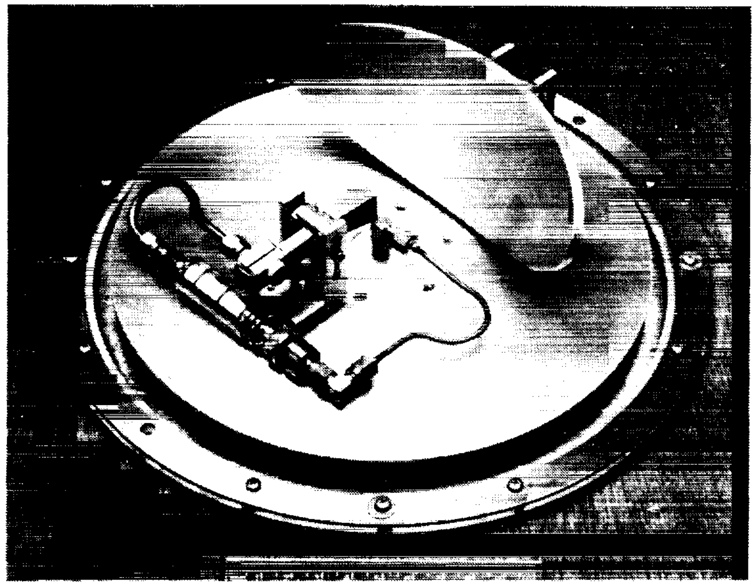


Figure 2. AMT Reflector Antenna and Test Platform

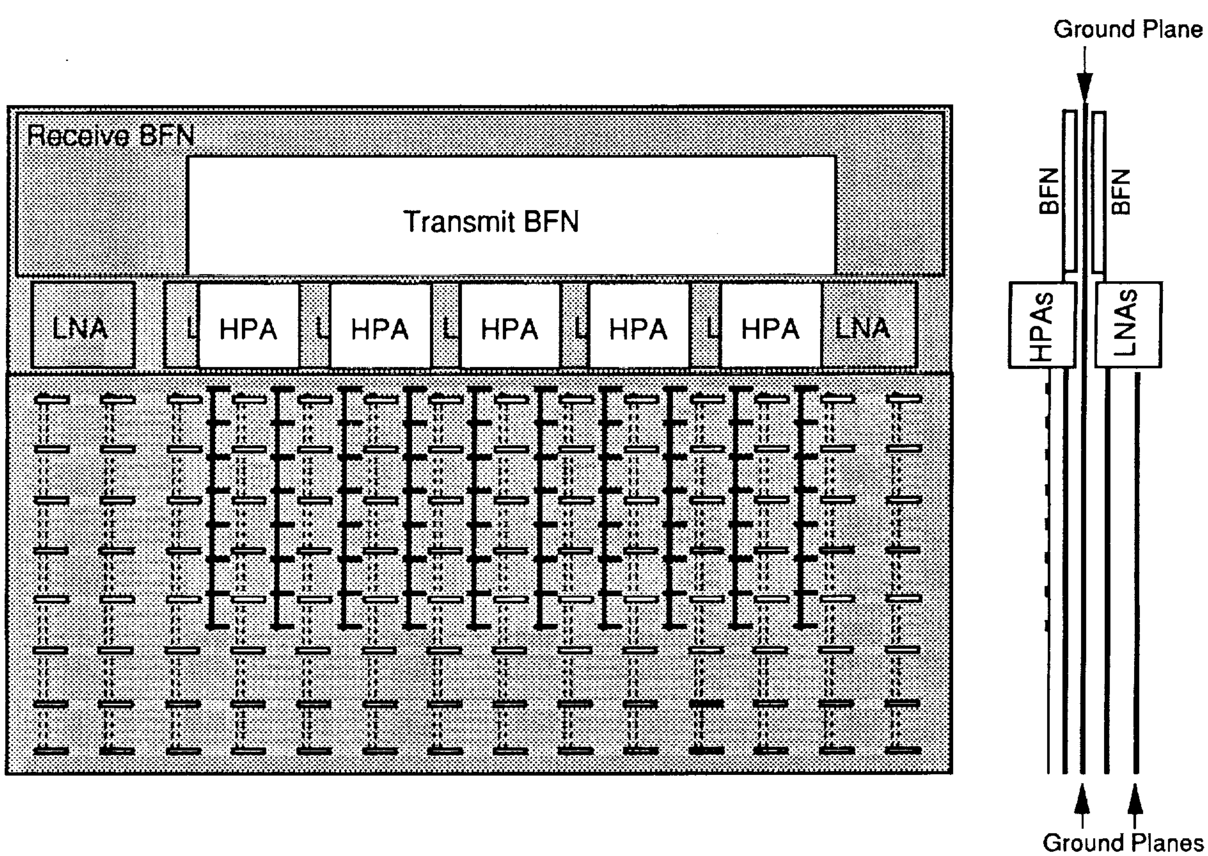


Figure 3. Concept drawing of the AMT Planar Active Array

N92-22433

ANTENNAS FOR MOBILE SATELLITE COMMUNICATIONS

John Huang
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California 91109

ABSTRACT

A NASA sponsored program, called the Mobile Satellite (MSAT) system, has prompted the development of several innovative antennas at L-band frequencies. In the space segment of the MSAT system, an efficient, light weight, circularly polarized microstrip array that uses linearly polarized elements has been developed as a multiple-beam reflector feed system. In the ground segment, a low-cost, low-profile, and very efficient microstrip Yagi array has been developed as a medium-gain mechanically steered vehicle antenna. Circularly shaped microstrip patches excited at higher-order modes were also developed as low-gain vehicle antennas. A more recent effort called for the development of a 20/30 GHz mobile terminal antenna for future-generation mobile satellite communications. To combat the high insertion loss encountered at 20/30 GHz, series-fed MMIC (Monolithic Microwave Integrated Circuit) microstrip array antennas are currently being developed. These MMIC arrays may lead to the development of several small but high-gain Ka-band antennas for the Personal Access Satellite Service (PASS) planned for the 2000s.

INTRODUCTION

The Mobile Satellite (MSAT) system is a satellite-based L-band communications network that is capable of providing telephone and data services to mobile users throughout a vast geographical area. Studies in the early 1980s by NASA, Canada, private sector, and others all point to the fact that the high capacity mobile satellite systems are viable only if certain high-risk enabling technologies are developed[1]. Currently, the Jet Propulsion Laboratory (JPL) has completed the managing of the development tasks[2] for the NASA sponsored MSAT program initiated in 1984. It is now the responsibility of the private industry to further develop, manufacture, own, and operate the MSAT as a commercial system.

One of the concentrated areas of technology development by JPL is antennas for satellite and land vehicles. In the space segment, since both the satellite and land vehicles are power limited, it is essential to develop a high-gain and efficient satellite antenna system. A circularly polarized microstrip array composed of simple linearly polarized elements[3] has been developed as an efficient, compact, and light-weight array feed for a high-gain multiple-beam reflector antenna system[4,5]. In the ground segment, the vehicle antenna not only needs to be affordable to users in price but also should be aerodynamically and aesthetically appealing. Both a mechanically steered medium-gain microstrip Yagi array[6] and a low-gain higher-order mode circular microstrip patch[7] have been developed as the low-profile and low-cost vehicle antennas.

For the NASA's Advanced Communication Technology Satellite (ACTS) to be launched in 1992, various experiments have been planned to demonstrate communication technologies at Ka-band frequencies. One of the AMT's (ACTS Mobile Terminal) high risk technologies that are currently being developed at JPL is a compact and low-cost mobile antenna system. This is a mechanically steered series-fed microstrip array using active MMIC components. This antenna development, along with other Ka-band technologies, will lead to the development of several compact hand-held terminals in a future Personal Access Satellite Services (PASS) system[13].

Both the above mentioned L-band and Ka-band satellite communication antenna developments and concepts are individually described in the following sections.

CIRCULARLY POLARIZED ARRAY COMPOSED OF LINEARLY POLARIZED ELEMENTS

Very large multiple-beam reflector antennas in the 20 to 50-meter range have been proposed for the MSAT satellite[1]. From 40 to 90 contiguous beams covering continental U.S. (CONUS) are to be generated from the reflector via an overlapping cluster feed array[4,5] with dimensions up to 6 meters. A structure of this size should have the capability of being folded and stowed in a space-limited satellite launching vehicle. As a consequence, the feed array should be low in profile and light in weight. Microstrip radiator was selected to meet these challenges. In order for the array to cover both the downlink frequencies (1545 to 1559 MHz) and the uplink frequencies (1646 to 1660 MHz), either a single patch with relatively thick substrate (0.5 inch) and 4 feeds or a dual-stacked patches with two feeds are needed to meet the bandwidth and circular polarization (CP) requirements. For a large array, such as relatively complicated element would increase the complexity of an already complex beam-forming feed network. Circularly polarized array composed of single-feed linearly polarized microstrip elements[3] has been developed to counter this complexity problem. The CP is achieved by having a basic 2x2 subarray, as shown in Figure 1, with elements' angular orientation and feed phases arranged in the sequential 0°, 90°, 180°, 270° manner. With such a system, not only is the feed complexity reduced, but also the bandwidth performance is improved. It has been found, however, that the basic 2x2 array does not generate acceptable CP away from principal planes (worst in the diagonal planes). Large cross-pol radiations (approximately -5 dB below main beam peak) are found in the off-broadside region due to amplitude and phase imbalances. This high cross-pol in the diagonal planes can be suppressed in a large array due to array factor's narrower beam and due to an averaging effect so that the imbalances are averaged out.

One of the arrays that have been fabricated and tested to demonstrate the above achievement is a 28-element microstrip array[5] as illustrated in Figures 2 and 3. It is only a single cluster array out of the many cluster arrays to be used in the multiple-beam reflector feed system. Amplitude taper is introduced in the array to suppress the grating lobes caused by large element spacings and to generate proper edge taper for the reflector. Each element is a single-probe fed half-inch thick honeycomb-supported square patch. Measured spinning-dipole patterns in the diagonal plane of the array are given in Figure 4 where excellent CP performance can be observed within the bandwidth (7.5%) of the uplink and downlink frequencies.

One important note on this antenna concept is that, even in a large array, if the element spacing becomes large (>0.55 wavelength) the antenna will start to lose gain[8] (up to 3 dB) due to the high cross-pol and grating lobes formed in the diagonal planes. Consequently, when designing this antenna concept, careful attention should be given to the element spacing. One other note is that this array concept of generating CP by using LP elements can be applied not only on microstrip patches but also on other types of radiators such as dipoles, horns, helixes, etc.

MICROSTRIP YAGI ARRAY

A major element of the MSAT program has been the development of several types of medium-gain L-band vehicle antennas. These antennas are required to generate CP with 10 dBic of minimum gain within the angular region of 20° to 60° above the horizon and shall be able to track the geostationary satellite while the vehicle is moving about in the CONUS. Previously, two types of medium-gain antennas have been developed and field tested with successful results. One is electronically steered planar phased array[9] and the other is the mechanically steered 1x4 tilted microstrip patch array[10]. The phased array offers the advantages of low-profile (1.0-inch thick) and beam agility at the expense of very high production cost (several thousand dollars per unit). On the other hand, the mechanically steered 1x4 array antenna offers a low production cost (several hundred dollars per unit) with a high profile (6-inch tall).

To combat the disadvantages of high cost and high profile of the above antennas, a new antenna concept, called the microstrip Yagi array[6], has been developed, which not only offers low profile and low cost but also shows excellent efficiency in its beamforming circuitry. This antenna system, as depicted in Figure 5, is a mechanically steered low-profile array composed of twelve parasitic director and reflector patch elements and four driven elements. The reflector and director patches, based on Yagi-Uda's principle, tilt the array's beam close to endfire for satellite pointing. Because only few driven elements are directly connected to the RF power distributing circuitry, the complexity and loss of this circuit are dramatically reduced, resulting in improved antenna noise temperature and antenna gain. The overall height of the integrated antenna, as shown in Figure 6, is only 1.5 inches, which includes both the RF portion and the mechanical rotating platform. Diameter of the radome is 21 inches. The rotation in azimuth is accomplished by the thin pancake motor. The control of this motor, or the beam pointing, is done by a monopulse system that is uniquely designed[11] for the communication antenna to provide simultaneous transmit and receive signals. Since the beamwidth in elevation is fairly wide, no elevation tracking of the satellite is necessary. To facilitate the ease of design and have a complete understanding of this microstrip Yagi array, a theoretically model based on the Method of Moments has been developed by the University of Massachusetts[12].

The microstrip Yagi array utilizes the same principle as a conventional dipole Yagi array where the electromagnetic energy is coupled from the driven element through space into the parasitic elements and then re-radiated to form a directional beam. For the microstrip Yagi, however, the adjacent patches need to be placed very close to each other so that significant amount of coupling can be formed through surface wave and radiation. Since the amount of surface wave is a strong function of the dielectric constant and substrate thickness, the pattern shape of the microstrip Yagi is also a function of these two parameters. Detailed antenna dimensions, description of the beamformer circuitry, and performance results are given in references 6 and 11. Good CP quality, adequate bandwidth, and excellent gain have been achieved by the microstrip Yagi array for the MSAT system. A typical elevation pattern showing both the calculated and measured results is presented in Figure 7. Since two major components of the antenna system, the radiating patches and the beamformer, can be manufactured by simple etching process, a relatively low-cost antenna system have been realized. The estimated manufacturing cost with a 10,000-unit per year and a 5-year production is about $450 per unit.

HIGHER-ORDER MODE CIRCULAR MICROSTRIP ANTENNA

An alternative approach to the MSAT medium-gain vehicle antenna is to increase the capacity of the spacecraft antenna while having a low-cost and low-gain vehicle antenna. When produced in mass quantity, this low-gain antenna is expected to cost only ten's of dollars and thus making the ground terminal more affordable to average consumers. This vehicle antenna should have a minimum gain of 3 to 4 dBic throughout the elevation angular region of 20° to 60° above the horizon. To provide such a coverage, a conical pattern (null at zenith) is preferred. Two antennas that have been thoroughly investigated for this application are the crossed drooping-dipoles and the quardrifilar helix. A third antenna, that has a low-profile and can be conformal to the car's rooftop, is the higher-order mode circular microstrip patch[7]. It is not only aesthetically appealing but also has a better chance to survive the abuses of a car wash.

When a circular patch has a dimension that is larger than the fundamental mode ($TM_{nm} = TM_{11}$) patch, higher-order mode (TM_{nl}) can be excited. The higher-order modes will produce conical patterns while the fundamental mode can only radiate a broadside beam. Two feeds with proper angular spacing and 90° phase differential are required to generate CP from a circular patch. Besides the desirable mode, there are generally many undesirable modes present in the patch cavity with less magnitudes. To preserve pattern symmetry and to keep cross-polarization low, especially for relatively thick substrate, the undesirable modes need to be suppressed. Generally, the two neighboring modes of a desirable resonant mode have the next highest magnitude. One way to suppress these adjacent modes is to employ two additional feeds located diametrically across from the two original feeds. Together, these four feeds, as illustrated in Figure 8, should have a phase arrangement of 0°, 90°, 0°, 90° for even-order modes and 0°, 90°, 180°, 270° for odd-order modes, so that the fields of the undesirable modes from the two opposing feeds cancel. It was found that the peak of the conical pattern can be changed over a wide angular range from 35° to about 60° from the disk broadside, depending on the substrate's dielectric constant and the resonant mode order. The higher the dielectric constant or the mode order, the larger the peak angle of the antenna is from the broadside.

Two circular microstrip antennas with CP have been constructed and tested. One was constructed on a honeycomb substrate (relative dielectric constant = 1.2) with TM_{21} mode excitation. It has a measured radiation peak at 36° from the broadside. The other was constructed on a fiberglass-reinforced teflon substrate (relative dielectric constant = 2.17) with TM_{41} mode excitation. It has a measured radiation peak at 55° from the broadside. Both antennas show good agreement between measured and calculated results as presented in Figure 9. The calculation was done by employing the Multimode Cavity Theory augmented with the Geometrical Theory of Diffraction[7].

Ka-BAND SMALL TERMINAL ARRAYS

The NASA/JPL AMT project is currently developing a mechanically steered Ka-band microstrip array with active MMIC components. To combat the high antenna insertion loss that incurred at Ka-band, the distributed MMIC high-power amplifiers (HPA) and low-noise amplifiers (LNA), along with a series-fed microstrip array technology, are being developed. The planar array, as shown in Figure 10, will be tilted and fixed in elevation and mechanically rotated in azimuth to track the satellite as the vehicle is in motion. It will have an elevation beamwidth of 12° and an azimuth beamwidth of 3° to 5° with a peak gain of 25 dBi. The 20 GHz receive array will be vertically polarized, while the 30 GHz transmit array is to be horizontally polarized. Total size of the antenna radome should be within 8-inch in diameter and 3-inch in height. The technical challenge in developing this antenna are: 1) to minimize the insertion loss in the series-fed microstrip array, 2) to achieve the required antenna gain within the given size limit, and 3) to package the MMIC components effectively.

The experience that is being learned from the AMT antenna development will lead to the development of several compact "hand-held" terminals that are being planned for the future 20/30 GHz PASS program. Several antenna concepts have been generated and are illustrated in Figures 11 through 15. The chief advantage of the head-mounted antennas is to minimize possible RF damage to the human eyes. Microstrip type of radiating elements along with compact and low-loss array feed technologies, as well as MMIC active devices, are again to be the areas of vital technology developments.

SUMMARY

Several innovative antenna techniques have been developed for the L-band MSAT system, as well as a number of antenna concepts are being developed or proposed for the Ka-band satellite communication systems. Due to the required small size and low cost for these antennas, microstrip printed antenna technologies have been the main thrust of the development effort. Several challenging areas in developing these antenna are the effective generation of circular polarization, insertion loss minimization, antenna size and manufacturing cost reduction, and effective packaging of the MMIC components.

ACKNOWLEDGEMENT

The research described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

REFERENCES

1. F. Naderi, et al, "NASA's Mobile Satellite Communications Program," 35th International Aeronautics Federation Congress, Laussane, Switzerland, October 10, 1984.

2. N. Lay, et al, "Description and Performance of a Digital Mobile Satellite Terminal," International Mobile Satellite Conference, Ottawa, Canada, PP. 272 - 278, June 1990.

3. J. Huang, "A Technique for an Array to Generate Circular Polarization with Linearly Polarized Elements," IEEE Trans. Antennas and Propag., Vol. AP-34, PP. 1113 - 1124, September 1986.

4. V. Jamnejad, "Multibeam Feed System Design Considerations .... ," NASA Conference on Large Space Antenna Technology, December 1984.

5. J. Huang and V. Jamnejad, "A Microstrip Array Feed for Land Mobile Satellite Reflector Antennas," IEEE Trans. Antennas and Propag., Vol. 37, PP. 153 - 158, February 1989.

6. J. Huang and A. Densmore, "Microstrip yagi Array Antenna for Mobile Satellite Vehicle Application," IEEE Trans. Antennas and Propag., Vol. 39, PP. 1024 - 1030, July 1991.

7. J. Huang, "Circularly Polarized Conical Patterns from Circular Microstrip Antennas," IEEE Trans. Antennas and Propag., Vol. AP-32, PP. 991 - 994, September 1984.

8. P. S. Hall, et al, "Gain of Circularly Polarized Arrays Composed of Linearly Polarized Elements," Electronics Letters, Page 124, January 1989.

9. J. Huang, "L-Band Phased Array Antennas for Mobile Satellite Communications," IEEE 37th Vehicular Technology Conference, Tampa, Florida, June 1987.

10. V. Jamnejad, "A Mechanically Steered Monopulse Tracking Antenna for PiFEX," JPL MSAT-X Quarterly Report No. 13, 1988.

11. J. Huang, et al, "Microstrip Yagi Array for MSAT Vehicle Antenna Application," International Mobile Satellite Conference, Ottawa, Canada, PP. 554 - 559, June 1990.

12. D. Pozar, "Design of a Microstrip Yagi Array for Application in the Mobile Satellite Program," Final Report for JPL Contract No. 958634, by University of Massachusetts, March 1990.

13. Sue, M.K., "Personal Access Satellite System Concept Study," JPL report D-5990, February 1989.

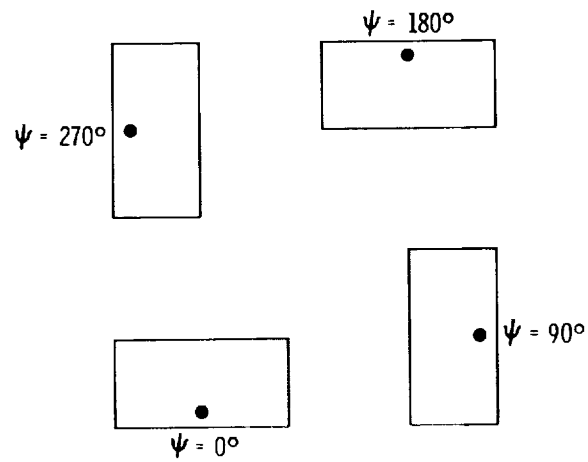


Figure 1. Basic 2x2 LP microstrip elements that generate CP.

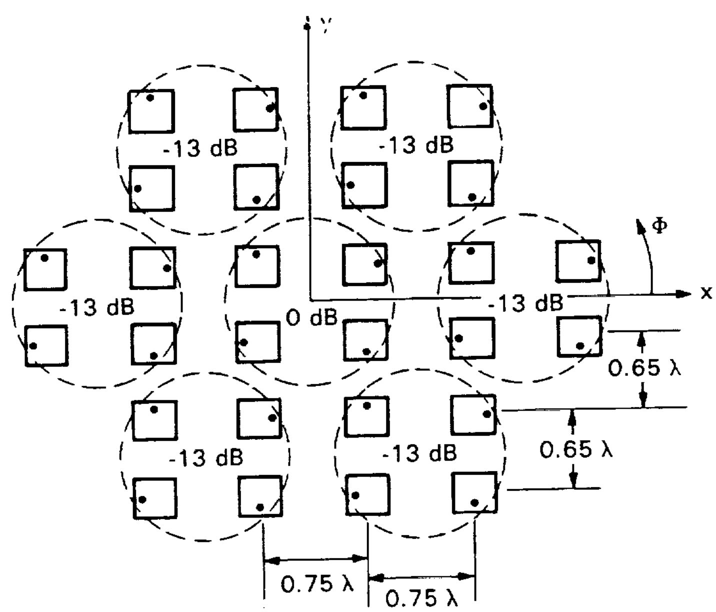


Figure 2. Single cluster 7-subarray microstrip array feed design.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

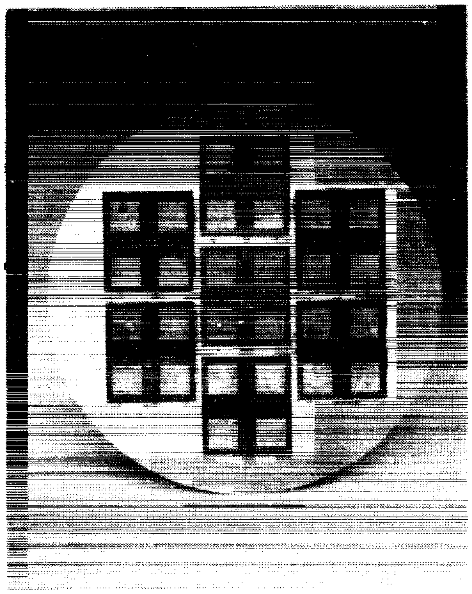


Figure 3. Constructed single cluster 7-subarray microstrip array feed.

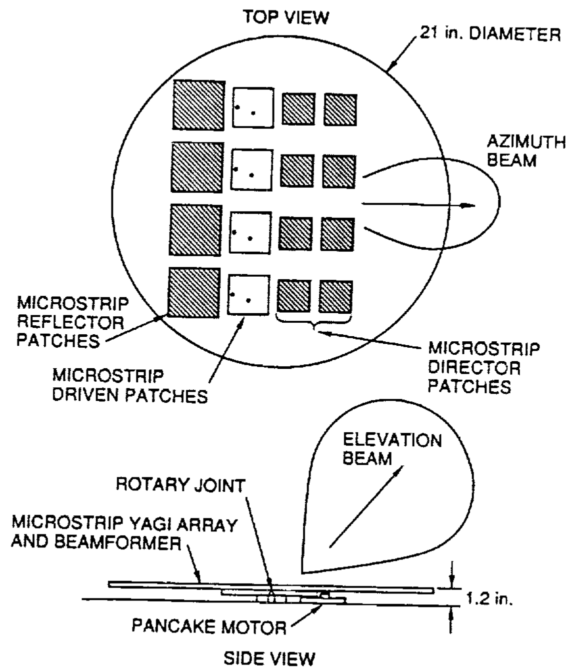


Figure 5. Configuration of MSAT microstrip Yagi array antenna.

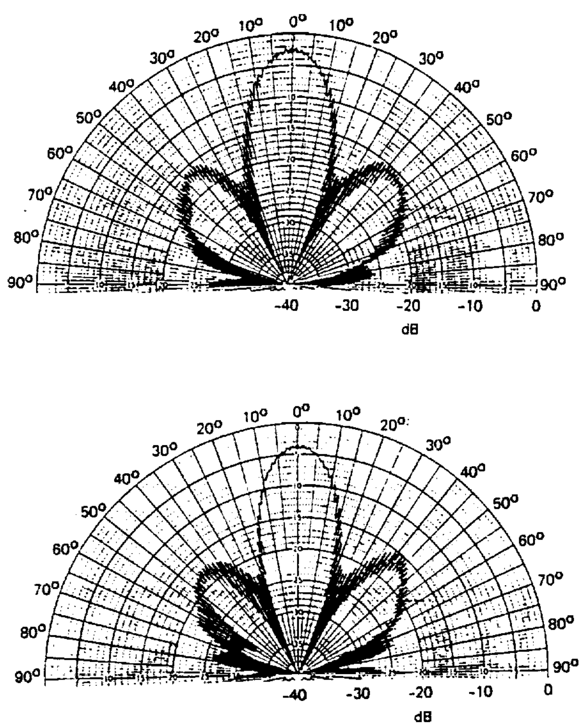


Figure 4. Spinning dipole patterns. Top 1.54 GHz, bottom 1.66 GHz.

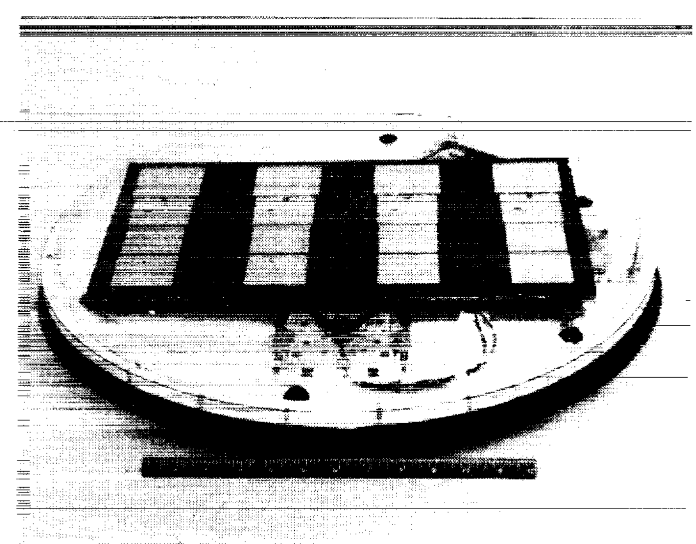


Figure 6. Integrated MSAT microstrip Yagi array antenna.

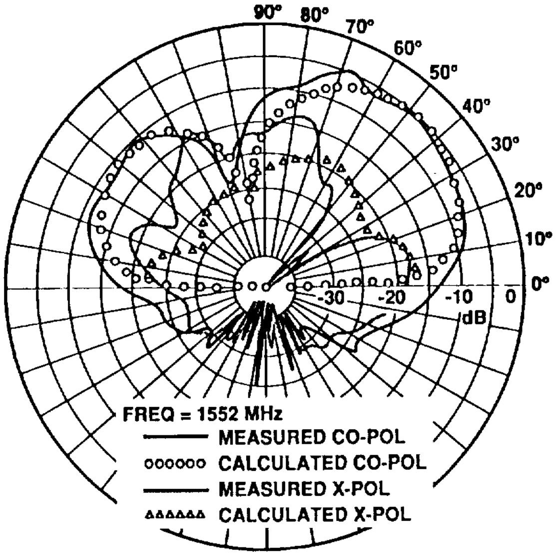


Figure 7. Elevation pattern of microstrip Yagi array.

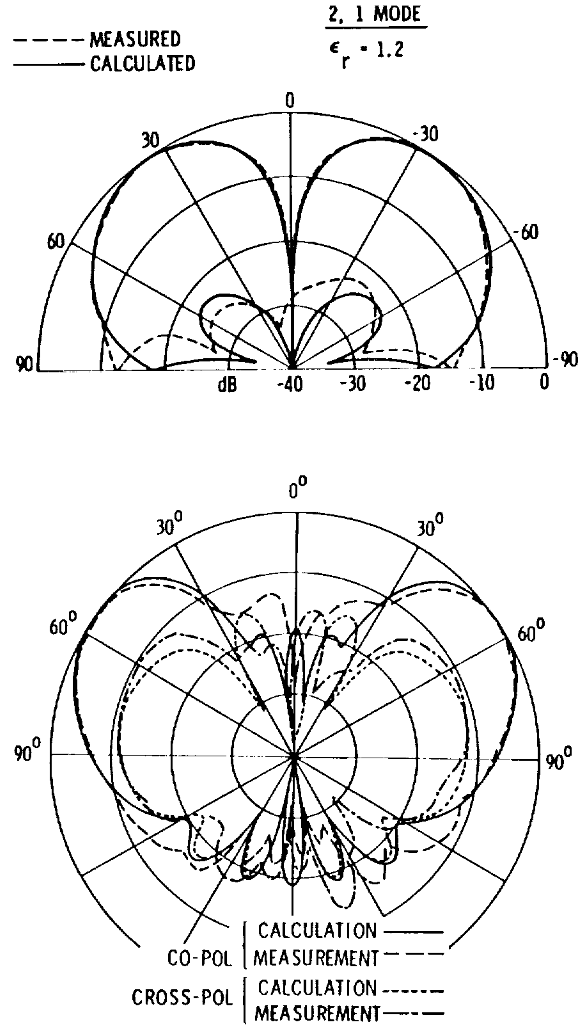


Figure 9. Patterns of circular microstrip patches with CP. (a) TM_{21} mode, (b) TM_{41} mode.

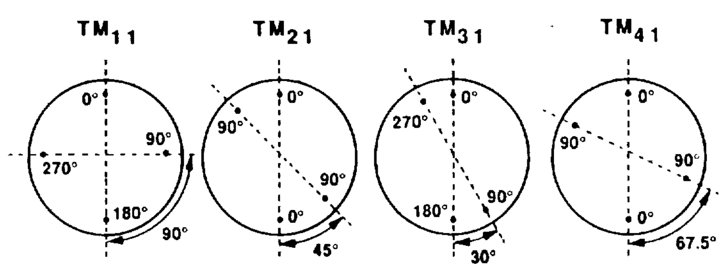


Figure 8. Four-probe feeds for circular patch at different resonant modes.

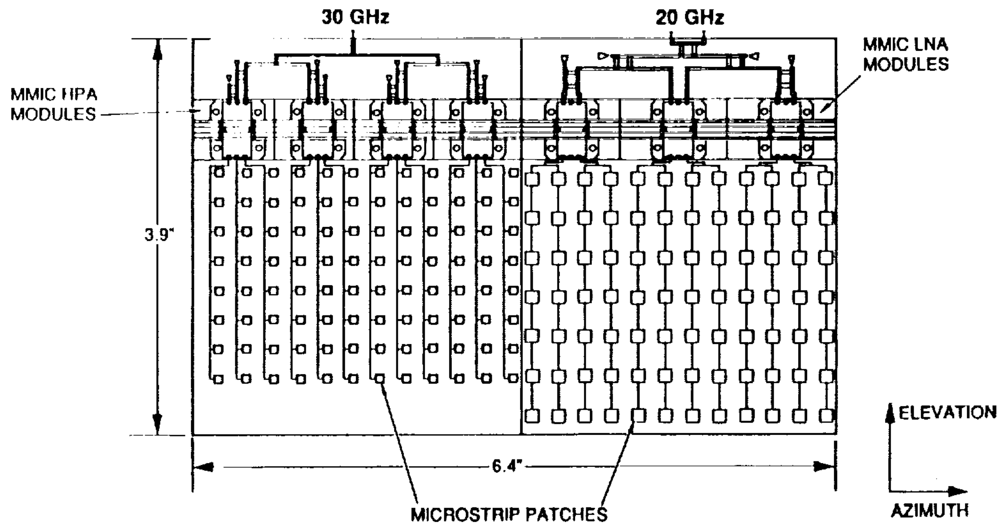


Figure 10. Ka-band microstrip array antenna with MMIC components.

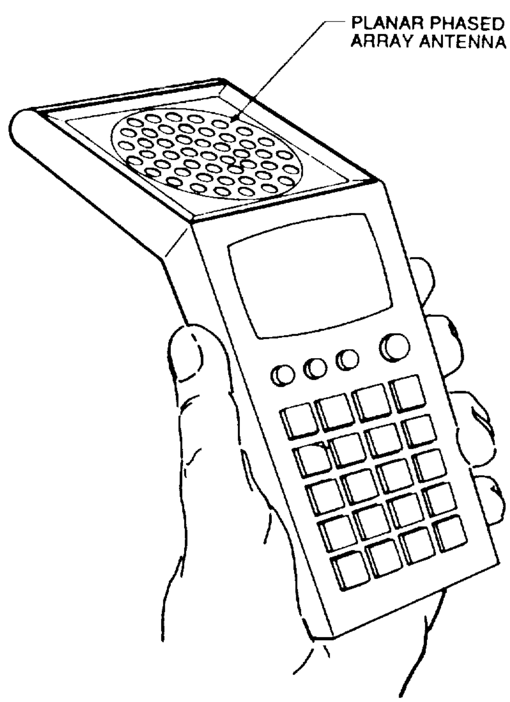


Figure 11. Hand-held terminal with MMIC microstrip phased array.

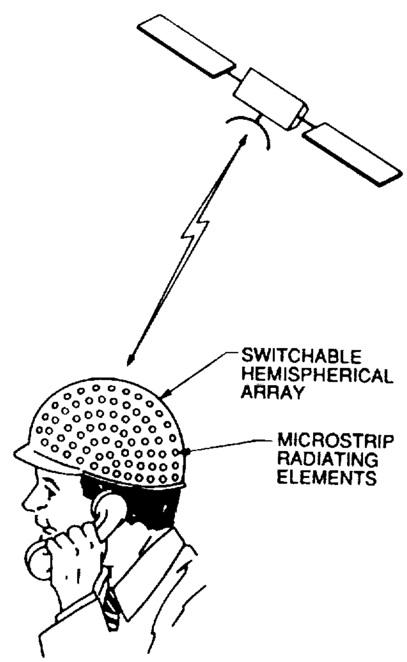


Figure 12. Hemispherical switching array.

Figure 13. Head-mounted array antenna.

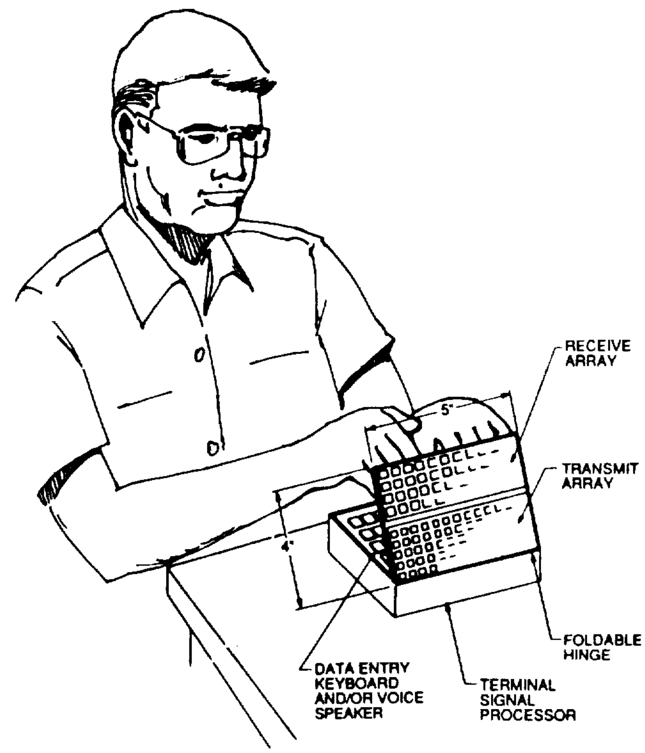


Figure 15. Lap-top or desk-top terminal with microstrip array antennas.

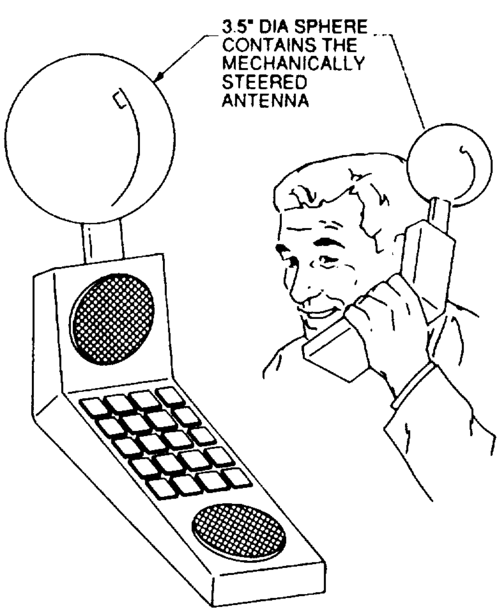


Figure 14. Hand-held terminal with mechanically steered antenna.

N92-22434

MMIC LINEAR-PHASE AND DIGITAL MODULATORS FOR DEEP SPACE SPACECRAFT X-BAND TRANSPONDER APPLICATIONS

Narayan R. Mysoor
Jet Propulsion Laboratory, California Institute of Technology
Pasadena, CA 91109

Fazal Ali
Pacific Monolithics
Sunnyvale, CA 94086

ABSTRACT

This article summarizes the design concepts, analyses and the development of GaAs monolithic microwave integrated circuit (MMIC) linear-phase and digital modulators for the next generation of space-borne communications systems. The design approach uses a very compact novel lumped element quadrature hybrid and MESFET-varactors to provide low loss and well-controlled phase performance for deep space transponder (DST) applications. The measured results of the MESFET-diode show a capacitance range of 2:1 under reverse bias, and a Q of 38 at 10 GHz. Three cascaded sections of hybrid-coupled reflection phase shifters have been modeled and simulations performed to provide an X-band (8415 ± 50 MHz) DST phase modulator with ± 2.5 radians of peak phase deviation. The modulator will accommodate downlink signal modulation with composite telemetry and ranging data, with a deviation linearity tolerance of ± 8% and insertion loss of less than 8 ± 0.5 dB. The MMIC digital modulator is designed to provide greater than 10 Mb/s of bi-phase modulation at X-band.

I. INTRODUCTION

GaAs MMIC analog linear-phase and digital modulators have been analyzed and investigated to provide the capability to directly modulate an X-band (8415 MHz) downlink carrier for deep space transponder (DST) applications [1]. The design specifications [1, 2] for the analog and digital phase modulators are given in Tables 1 and 2, respectively. The analog phase modulator [2] must be capable of large linear phase deviation, low loss, and wideband operation with good thermal stability. In addition, the phase modulator and its driver circuit must be compact and consume low dc power. The design is to provide ± 2.5 radians of peak phase deviation to accommodate downlink modulation of telemetry and ranging signals. The tolerance on the phase deviation linearity is ± 8%. The insertion loss should be less than 8 dB and its variation with phase shift should be within ± 0.5 dB. The phase delay variation specifications over the transponder hardware qualification environment, - 30°C to + 85°C is less than 0.5 ps/°C for the phase modulator. This investigation will consider the reflection type phase shifter for the GaAs MMIC implementation of the hardware. The digital modulator is designed to provide a binary phase shift Keying (BPSK) modulation of 10 Mb/s. The organization of the article is as follows. The modulator design and circuit configurations are presented in Section II. The test data is presented in Section III. The conclusions are presented in Section IV.

II. GaAs MMIC LINEAR-PHASE AND BI-PHASE MODULATORS

A. GaAs MMIC Linear-Phase Modulator Design

The X-band MMIC linear-phase modulator is based on a reflection type hybrid coupled phase shifter design approach [3, 4]. A basic building block of a single-section continuously variable reflection phase shifter is shown in Fig. 1. It consists of a pair of MESFET-varactors coupled to a lumped element quadrature coupler [5] with a series inductor and a parallel resistor. The X-band lumped element 3 dB hybrid coupler consists of a multi-turn bifilar spiral transformer and capacitors. The design equations for this lumped element coupler are given in the references, [5] and [6]. The capacitors were implemented using a standard metal-insulator-metal (MIM) structure with silicon nitride as the dielectric insulator material. The MESFET used in this application is a standard 0.5X600 micron depletion mode MESFET from Triquint analog MMIC process. The gate length of the MESFET is equal to half micron. The MESFET-varactors are designed by shorting the source and drain electrodes to form the cathode. The MESFET-varactors are readily integrable with inductors, capacitors, and resistors needed for the rest of the MMIC circuit. The model of the MESFET-varactors used in the CAD simulations is based on the measured capacitance-voltage (C-V) characteristics over 0.1 to 10 GHz frequency range. The measured C-V characteristics showed greater than 2:1 capacitance change over the bias range of -2 V to -6 V. The MESFET-varactor quality factor (Q) is approximately equal to 38, and the estimated cut-off frequency is 380 GHz. The MESFET-varactor, series inductor L1 and lumped hybrid are optimized to provide a linear phase shift of 100 degrees. A resistor is shunted across each diode to minimize insertion loss variation as the phase is varied.

An X-band ± 2.6 radians (300 degree) analog phase shifter was designed and fabricated using a cascade of three single-section hybrid coupled phase shifter circuits as shown in Fig. 2. Buffer amplifiers are used to provide an isolation of about 38 dB between sections. The buffer amplifiers use 0.5X200 micron MESFETs in a cascode amplifier circuit configuration. This linear-phase modulator [Fig. 2] has been simulated using CAD tools to have ± 2.6 radians of continuous phase shift in the 8.4 to 8.6 GHz range with the variation of the MESFET-varactor control voltage from -2 V to -6 V. The simulated insertion loss is 0 dB, with ± 0.5 dB variation over the control voltage. The input and output return losses were better than 15 dB in simulation.

The X-band linear-phase modulator [Fig. 2] chip measures 96X36 mils. This compact layout using lumped elements is about one-quarter the size of the conventional distributed element design approach.

B. GaAs MMIC Bi-Phase Modulator Design

The block diagram of the MMIC bi-phase modulator is shown in Fig. 3. It consists of a lumped element quadrature hybrid and MESFET switches. The design is based on the reflection phase shifter approach [3 - 5]. When the MESFET switches are "on", the phase modulator is in the minimum phase state and vice versa. By switching the MESFETs from "on" to "off" state, one can theoretically achieve 180 degrees differential phase shift in the 8 GHz to 9 GHz band. Both input and output return losses are better than 15 dB at both states. The insertion loss variation between the two states was simulated as ± 0.2 dB.

The bi-phase modulator design resulted in a small chip size of 36X36 mils. To the best of our knowledge, this is the smallest chip size reported for an X-band BPSK modulator.

III. MEASURED RESULTS

The measured phase shift for the X-band linear-phase modulator [Fig. 2] is shown in Fig. 4. The phase measured at 2 V reverse bias was used as the reference. From 2 V to 6 V reverse bias, a linear phase shift of ± 2.6 radians (300 degrees) was obtained between 8.2 to 8.7 GHz. The phase shift as a function of the bias voltage is shown in Fig. 5 for frequencies 8.4 GHz and 8.5 GHz. The phase shift linearity is better than ± 2% over the frequency band of interest for DST application; 8.4 GHz to 8.5 GHz. For reverse bias, the measured insertion loss over this frequency band was - 2 ± 1.5 dB. The insertion loss variation needs to be reduced to ± 0.5 dB to satisfy the specified requirements [Table 1]. A second design/fabrication iteration has been planned. The device models will be revised to fit the experimental data, and the circuit components will be optimized to meet the specifications in the next iteration.

The measured phase shift performance and insertion loss for the bi-phase modulator are shown in Figs. 6 and 7, respectively. The 180 degrees phase shift has a maximum of ± 4 degrees of phase error over the frequency range 8.2 GHz to 8.7 GHz. The measured phase shift is 180 ± 2 degrees, and insertion loss is 7 dB ± 0.3 dB over the DST downlink frequency range from 8.4 GHz to 8.5 GHz. As shown in Figs. 8 and 9, the input port return loss for the bi-phase is greater than 13 dB, and the output port return loss is greater than 10 dB. The bi-phase modulator does not meet the return loss requirement of 14 dB or better [Table 2] at its input and output ports. It will be optimized to meet this requirement in the next iteration.

IV. CONCLUSIONS

The development and performance X-band GaAs MMIC linear-phase and bi-phase modulators for the deep space transponder and space-borne communications systems applications are presented. The linear-phase modulator design is based on the reflection phase shifter approach and utilizes a novel lumped element quadrature hybrid and MESFET-varactors for each section. Three such sections have been used in tandem to achieve a linear phase shift in excess of ± 2.6 radians over 8.2 to 8.7 GHz frequency range. The insertion loss variation over the phase change is - 2 ± 1.5 dB over this frequency range, and it does not meet the desired specification. The shunt resistor across the MESFET-varactors will be adjusted and the circuit will be optimized in the second iteration to provide ± 0.5 dB insertion loss variation over the phase change. The chip size of the fabricated linear-phase modulator is 36X96 mils.

The GaAs MMIC bi-phase modulator is also based on the reflection phase shifter approach, and is composed of a lumped element quadrature hybrid and MESFET switches. The bi-phase modulator chip measures only 36X36 mils. The measured phase shift performance is 180 ± 2 degrees, and the insertion loss is 7 dB ± 0.3 dB over the frequency band from 8.4 to 8.5 GHz.

Potential commercial applications include phased arrays, and satellite communication systems. Commercial microwave systems which require continuous phase control in trimming multiple channels also benefit from this design.

ACKNOWLEDGMENTS

The research described in this paper was performed by Pacific Monolithics Co., under contract with the National Aeronautics and Space Administration Small Business Innovation Research (NASA SBIR) program. This research is done under the direction of NASA SBIR center Jet Propulsion Laboratory, California Institute of Technology. The authors gratefully acknowledge the valuable support by A. W. Kermode of Jet Propulsion Laboratory for comments on design.

REFERENCES

[1] N. R. Mysoor, J. D. Perret, and A. W. Kermode, "Design Concepts and Performance of NASA X-Band (7162 MHz/8415 MHz) Transponder for Deep-Space Spacecraft Applications," TDA Progress Report 42-104, vol. October-December 1990, Jet Propulsion Laboratory, Pasadena, California, pp. 247-256, February 15, 1991.

[2] N. R. Mysoor, and R. O. Mueller, "Design and Analysis of Low-Loss Linear Analog Phase Modulator for Deep Space Spacecraft X-band Transponder (DST) Application," TDA Progress Report 42-105, vol. January-March 1991, Jet Propulsion Laboratory, Pasadena, California, pp. 136-145, May 15, 1991.

[3] R. Garver, "360° Varactor Linear Phase Modulator," IEEE Trans. Microwave Theory Tech., Vol. MTT-17, pp 137-147, March 1969.

[4] C. L. Chen, W. E. Courtney, L. J. Mahoney, M. J. Manfra, A. Chu, and H. A. Atwater, "A Low-Loss Ku-Band Monolithic Analog Phase Shifter," IEEE Trans. on Microwave Theory and Tech., vol. MTT-35, no. 3, pp. 315-320, March 1987.

[5] F. Ali, and A. Podell, "Design and Applications of a 3:1 Bandwidth GaAs Monolithic Spiral Quadrature Hybrid," IEEE GaAs IC Symposium Digest, pp. 279-282, October 1990.

[6] F. Ali, and A. Podell, "A Wide-Band GaAs Monolithic Spiral Quadrature Hybrid and its Circuit Applications," IEEE J. Solid State Circuits, pp. 1394-1398, October 1991.

Table 1. MMIC Linear-Phase Modulator Specifications

PARAMETERS	SPECIFICATIONS
1. RF FREQUENCY	8415 ± 50 MHz
2. INSERTION LOSS	8 dB (MAX)
3. LOSS VARIATION	± 0.5 dB (MAX)
4. INPUT RETURN LOSS	14 dB (MIN)
5. OUTPUT RETURN LOSS	14 dB (MIN)
6. LINEAR PHASE SHIFT	± 2.5 RAD PEAK (285.5°)
7. MODULATION PHASE LINEARITY	± 5% BSL TO ± 2 RADIANS PEAK ± 8% BSL TO ± 2.5 RADIANS PEAK
8. MODULATION SENSITIVITY	2 RADIANS PEAK/VOLT PEAK
9. RF INPUT LEVEL	+ 10 dBm (MAX)
10. DC POWER	± 6V, 120 mW
11. DESIGN TEMP. RANGE	- 30°C TO + 85°C

Table 2. MMIC Bi-Phase Modulator Specifications

PARAMETERS	SPECIFICATIONS
1. RF FREQUENCY	8415 ± 50 MHz
2. INSERTION LOSS	7 dB (MAX)
3. LOSS VARIATION (AM)	± 0.5 dB (MAX)
4. INPUT RETURN LOSS	14 dB (MIN)
5. OUTPUT RETURN LOSS	14 dB (MIN)
6. RF INPUT LEVEL	+ 10 dBm (MAX)
7. PHASE SHIFT	180° ± 5°
8. BI-PHASE MODULATION	> 10 Mb/s
9. DESIGN TEMP RANGE	- 30° to + 85°C

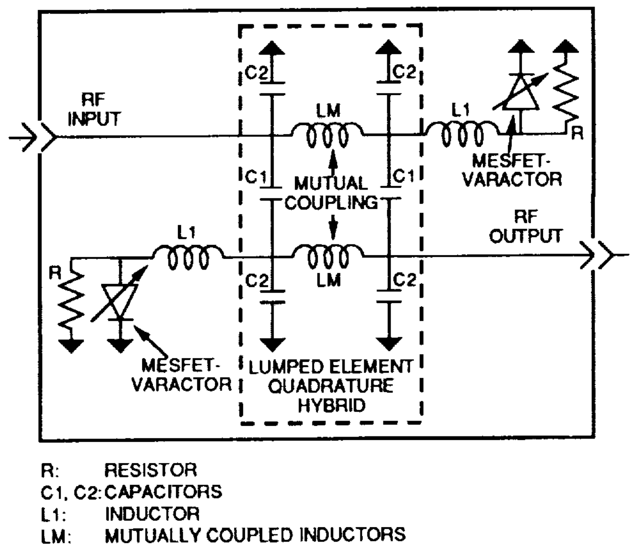


Figure 1. Single-section lumped element quadrature hybrid coupled phase shifter with MESFET varactors.

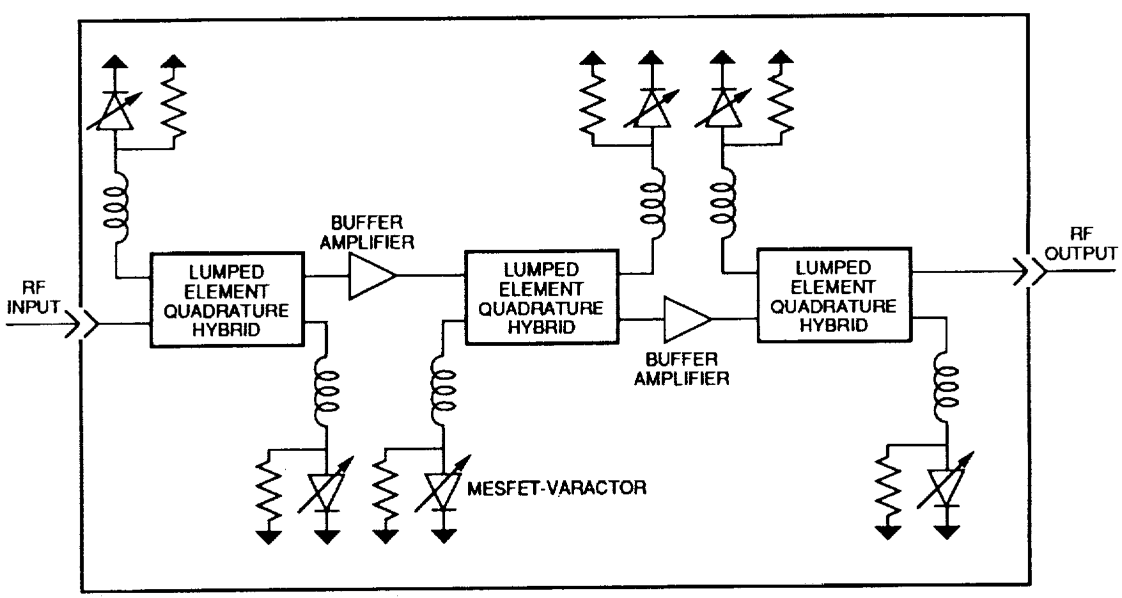


Figure 2. Block diagram of the three-stage linear phase modulator with two isolation amplifiers.

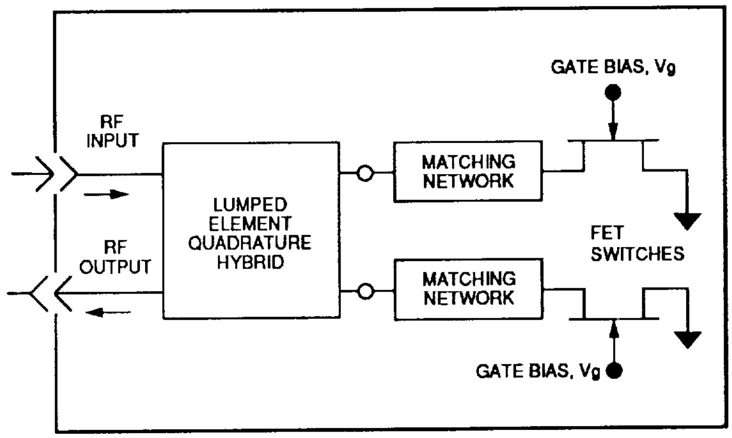


Figure 3. Block diagram of the bi-phase modulator with lumped quadrature hybrid.

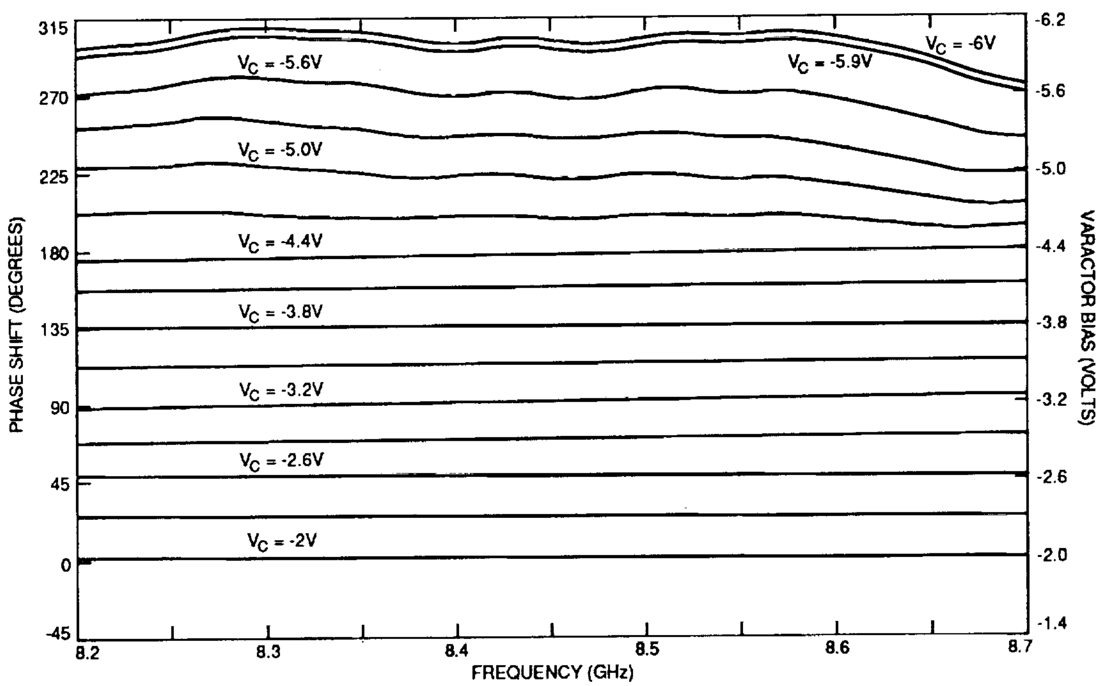


Figure 4. Measured phase shift as a function of the frequency at different dc bias levels. V_C is the dc bias control voltage.

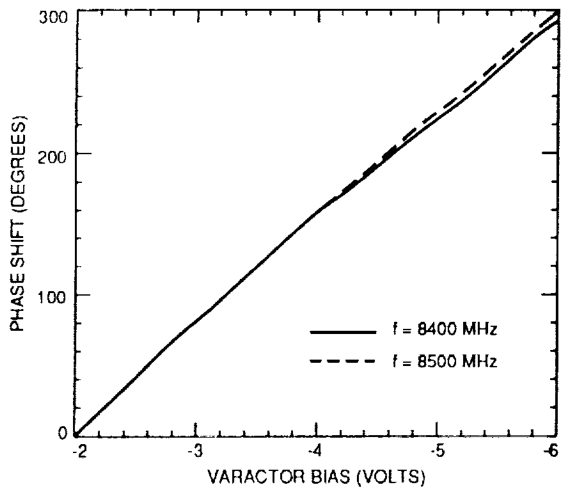


Figure 5. Measured phase shift as a function of the dc bias voltage at frequencies 8400 MHz and 8500 MHz.

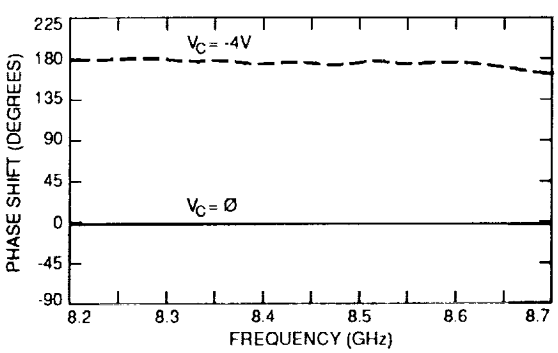


Figure 6. Measured phase shift performance of the bi-phase modulator as a function of the frequency.

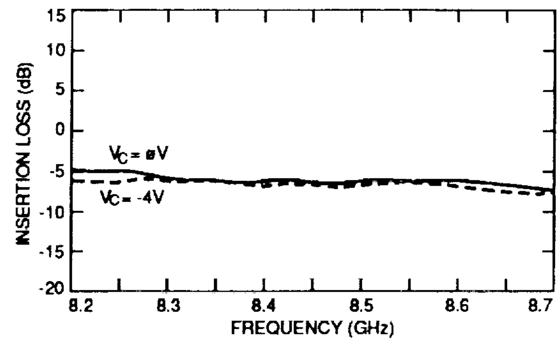


Figure 7. Measured insertion loss of the bi-phase modulator as a function of the frequency.

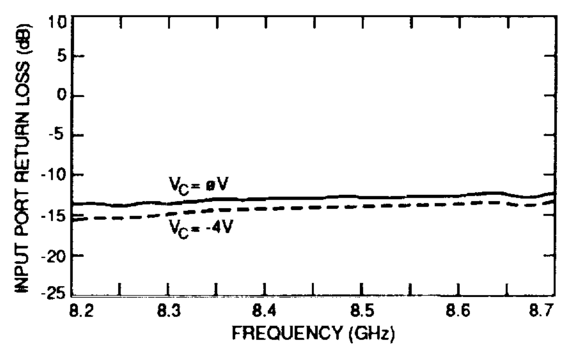


Figure 8. Measured input port return loss of the bi-phase modulator as a function of the frequency.

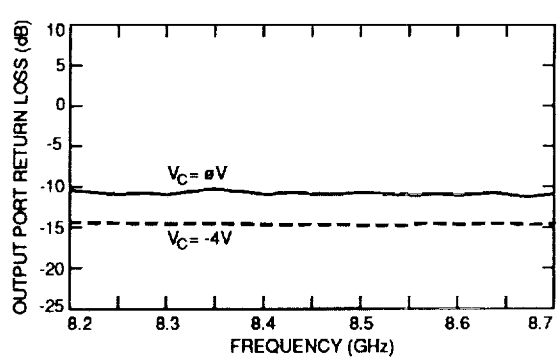


Figure 9. Measured output port return loss of the bi-phase modulator as a function of the frequency.

N92-22435

PHASED ARRAY ANTENNA BEAMFORMING USING OPTICAL PROCESSOR

L. P. Anderson, F. Boldissar, D. C. D. Chang
Hughes Aircraft Company, Space and Communications Group
P. O. Box 92919, El Segundo, CA 90245

ABSTRACT

This work concerns itself with the analytical investigation into the feasibility of optical processor based beamforming for microwave array antennas. The primary focus is on systems utilizing the 20/30 GHz communications band and a transmit configuration exclusively to serve this band.

A mathematical model is developed for computation of candidate design configurations. The model is capable of determination of the necessary design parameters required for spatial aspects of the microwave "footprint" (beam) formation. Computed example beams transmitted from geosynchronous orbit are presented to demonstrate network capabilities. The effect of the processor on the output microwave signal to noise quality at the antenna interface is also considered.

1. INTRODUCTION

Phased array antennas are playing an increasingly important role in current and anticipated radar and communications applications. RF distribution manifolds used to furnish the required aperture excitation can become prohibitively bulky for space applications. There are also limitations on microwave bandwidth and speed of beam scanning/reconfigurability with conventional methods.

Problems with conventional beamforming for large arrays makes it imperative to develop a proposed technology that performs array beamforming via alternative approaches. A novel method utilizing an optical signal processor to furnish these excitations has been proposed.

Keopf [1] has described the basic concept for both transmit and receive applications. In his description, a heterodyne processing scheme is used in conjunction with a static means (i.e. pinhole mask) for production of a scaled image of the desired far field pattern.

The scope of this effort was to derive a transmit architecture which would be tractable to formation of generalized shaped beam far field patterns. Thus a dynamic means for image formation must be employed. Additionally, the system should be capable of operation in the 20/30 GHz communications band. These requirements necessitated consideration of alternative approaches due to inherent limitations of the Keopf method in addressing these requirements. Once this basic processor configuration was established, a computer model of the network was generated for analysis of its beamforming capabilities.

A conceptual transmit mode processor is depicted in Figure 1. A continuous wave laser source emits a beam that is split into two paths. The optical signal in Path A undergoes temporal modulation with a microwave information signal. The optical signal in Path B is spatially modulated by a reflectance function, then Fourier transformed with a lens. The modulated optical signals from both of these paths are then added together and sampled spatially by a fiber optic array which transports the combined optical signal to the individual microwave array elements. At the array element the optical signal is detected to reproduce the microwave information signal. The spatial amplitude and phase distribution provided by Path B will be retained. This microwave signal is then amplified and radiated by the array antenna.

2. SPATIAL PROCESSING MODEL

The model (Figure 2) for the optical processor is based on the Fourier transform relationship between the front and rear focal planes of the optical lens of focal length f [2]. Here a two dimensional image formed in the rear focal plane can be expressible as an electric field distribution in the forward focal plane. The input field distribution is expressed as:

$$E(x_1, y_1) = t_0(x_1, y_1) E_0 \qquad (1)$$

where $t_0(x_1, y_1) \equiv$ image transmittance function of the SLM

the field formed in the front focal plane of the lens via diffraction are thus expressible through a Fourier transform i.e.,

$$E(x_2, y_2) \sim \iint_{-\infty}^{+\infty} t_0(x_1, y_1) \exp\left\{-\frac{j2\pi}{\lambda f}(x_1 x_2 + y_1 y_2)\right\} dx_1 dy_1 \qquad (2)$$

The pupil function is assumed to be unity for the lens. A digitized image of the desired far field footprint is provided in the form of light and dark contrasting elements (pixels). A discreet Fourier transform is thus formed for a NxM matrix scene of area axa (square) constant transmittance pixels. Thus the front focal field are now expressible by the double summation:

$$E_{TOTAL}(x_2, y_2) \sim \sum_{n=1}^{N} \sum_{m=1}^{M} t_{nm}(x_1, y_1) \operatorname{sinc}(a f_{x_n}) \operatorname{sinc}(a f_{y_m}) e^{j2\pi(b_n f_x + c_m f_y)} \qquad (3)$$

where f_x, f_y are spatial frequencies and bm, cn are offsets of the m, nth pixel in the x, y, directions respectively. The function

$$\operatorname{sinc}(z) \equiv \frac{\sin(\pi z)}{\pi z}$$

Thus both amplitude and phase (i.e. pixel "brightness" and location relative to a central axis) is contained in equation (3). The distribution E (x2,y2) now represents the complex aperture excitation, once downconverted, of the array antenna. The far zone fields resulting from such excitations may be obtained via standard Fraunhoffer diffraction evaluation of the array aperture.

Deterministic errors (i.e. axial defocussing of F.T.L.) can be approximated by the addition of a quadratic phase term to the incident wavefront i.e. [3]

$$E_{nm}(x_1, y_1) = E_0\left(t_{nm}(x_1, y_1) e^{-j\delta_{nm}}\right) \qquad (4)$$

Here the quadratic phase term is defined as:

$$\delta_{nm}(x_1, y_1) = k\left\{\frac{x_{1n}^2 + y_{1m}^2}{2R_0}\right\} \qquad (5)$$

where k is the free space wavenumber and Ro is the axial displacement of pixel from image plane. The effect of the error was evaluated using a one dimensional model to simplify understanding of the resultant Fourier transform spectra. For progressive values of phase error, the presence of grating lobes of increasing intensity is observed in the spatial frequency output plane. The element pattern presented by the individual pixels in the SLM serve to modulate the amplitude of the grating lobe formation. The existence of these lobes in the complex aperture distribution of the antenna array will serve to create secondary beams in the far field. Albeit in most applications, these lobes will not exist in visible space due to finite array aperture; they will tend to decrease main lobe directivity.

3. TEMPORAL PROCESSING MODEL

An essential element of consideration in the utilization of the processor is the evaluation of signal quality in the communications link. The underlying basis of the analysis is the development of the expression for processor output signal to noise ratio. It can be shown [4] that exclusive of input and output amplifiers and their associated matching circuits, the expression for processor (S/N) out is found to be

$$\left(\frac{S}{N}\right)_{out} = \frac{P_{in}\left(\frac{\pi P_n \eta_d e}{2 V_\pi L_o h\nu}\right)^2 R_D R_M}{\left(\frac{P_n \eta_d e}{2 L_o h\nu}\right)^2 R_D (RIN)\Delta\nu + \frac{P_n \eta_d e^2}{h\nu L_o}\left(1 - \frac{\eta_d}{2 L_o}\right) R_o \Delta\nu + kT\Delta\nu} \quad (6)$$

here the variables are defined as:

P_{in}	=	input RF signal drive power
P_n	=	laser optical power at nth array element
N_e	=	total number of microwave antenna array elements
R_D, R_M	=	resistance of photodiode and modulator respectively
V_p	=	voltage required for 180∞ phase shift in E/O modulator
Δn	=	operating RF bandwidth
n	=	operating optical frequency
h_o	=	photodiode efficiency
L_o	=	total OPBFN optical losses
RIN	=	laser random intensity noise
h,K	=	Planck's and Boltzman constants (respectively)
e	=	electronic charge

The noise contributions in the denominator can be attributed to three major categories [5]; (1) Laser noise, (2) partition noise (quantum effects) and (3) detector noise. Note that this general equation will be weighted accordingly with the individual amplitude coefficients associated with the F.T. spectra at the fiber optic input. Thus for uniform illumination:

$$P_T = N_e P_n \quad (7)$$

One can recognize the numerator as simply expressing the link gain i.e.;

$$G_{OPBFN} = \frac{P_{RFOUT}}{P_{RFIN}} = \left[\frac{\pi P_n \eta_d e}{2 V_\pi L_o h\nu}\right]^2 R_D R_M \quad (8)$$

A key aspect of the indirect modulation scheme in the OPBFN is that kink gain can be controlled by input laser power. This condition allows OPBFN link gain of unity or greater for certain network configurations. This advantage is not realizable in direct modulation schemes for OPBFN's. Because of CW laser operation in the former, the system performance does not degrade at high frequencies.

Utilization of equation (6) allows the study of (S/N) out as a function of input laser power. This can be expressed with the number of phases array elements as a parameter.

Figure 3 depicts this relationship. Here we have compared the effect of operational RF bandwidth on required optical power. We have observed that in general, the required optical power for a given array size and desired (S/N) out is proportional to the square root of the bandwidth. Thus the smaller the desired bandwidth, the less optical power is needed. As can be seen, twice as much optical power is required to meet minimum (S/N) out when the bandwidth is raised from 250 MHz to 1 GHz.

4. RESULTS

A theoretical model of a liquid crystal light valve (i.e. LCLV) was developed to study the aspects of beam formation using an optical processor based network. Figure 4 depicts a digitized image of a typical CONUS (Continental United States) footprint and selected regional coverages*. The chosen trial antenna aperture was a 121 element planar array, operating at a transmit frequency of 20 GHz. The total available aperture is 6273 square wavelengths. The particular grid density of 51 x 23 pixels in the LCLV array was chosen on the basis of the most efficient utilization of land mass coverage from the designated geosynchronous orbital position. Figures 5a, 5b, and 5c illustrate the versatility of the network by the creation of area coverage beams as well as multiple spot beams.

It might be noted that these simulations were done assuming an ideal LCLV with pixel transmittance being either 1 or 0 (i.e. infinite contrast ratio). In reality, there is always some leakage through the light valve so that the transmittance cannot go to zero. The effect of such leakage tends to broaden the main lobe and heighten the sidelobe levels of the microwave far field pattern. A parametric study was thus conducted to determine the lower limit of contrast ratio for adequate far field patterns to result. Criteria for evaluation were sidelobe level an main lobe beam width. It was determined that for the network under study, a minimum contrast ratio of 30 dB was required for spot beams and a 25 dB level for sector beams before a 10% increase in sidelobe level or half power beam width (as applicable) was observed.

A simple experiment was conducted to examine the validity of the theoretical spatial path model. Details of the complete evaluation can be found in [4]. A breadboard of the spatial path (i.e. Path B) was constructed on an optical bench and operated at 0.63 m using a standard He-Ne laser as the source illumination. An "off the shelf" Hughes model H01460 liquid crystal light valve was used to create images of full CONUS and quarter CONUS far field footprints. The output spatial frequency spectra (i.e. Fourier transform distribution) was used as the basis of comparison between theory and experiment. The model was adjusted to simulate a 20 dB background leakage in the LCLV image plane.

Initially photographs were taken to examine the qualitative nature of the optical transform spectra as compared to these computer predicted counterparts. The central lobe and first order sidelobe size and positions corresponded exactly with the predicted patterns once the photos were scaled. The scaling was necessary since the transform spectra were magnified through microscope objectives for evaluation of structural detail.

In order to check the quantitative nature of the computer predicted FT spectra at the processor output, measurement of electric field intensity of pattern cuts through both CONUS and quarter CONUS transforms were conducted. This was accomplished by using the digitizing CCD camera in conjunction with an oscilloscope for real time viewing of the voltage amplitudes. Horizontal pattern cuts were taken through both F.T. spectral outputs and exhibited reasonable agreement with computed intensities over the main lobe and first sidelobe regions. Figures 6 and 7 depict the comparison between computed and a measured output of the CCD camera for CONUS and quarter CONUS transform spectra respectively. As can be seen, dynamic range problems due to central lobe saturation of the camera prevented detailed pattern measurement beyond the second order lobe.

*Note that the image formed on the LCLV is an inverted scaled image of the desired far field pattern.

5. CONCLUSION

The concept of optical processor based microwave antenna beamforming has been deemed feasible supported by the analysis presented herein. Fundamental concepts of the theoretical model have been validated via experiment. Advantages and disadvantages of the approach as referenced to more conventional designs have been highlighted to establish the validity of this technique.

A quantitative view of the processor characteristics at the output of the antenna was presented. As a result, a generalized computer model has been established to aid in the processor design procedure.

Refinement of the analytical model and test of a breadboard transmit processor comprises the bulk of future work to be done. This includes the determination of processor configurations for receive applications. The breadboard model shall and in the understanding of single sideband modulation requirements of the processor as well as path length sensitivity.

REFERENCES

1. Keopf. G. A., "Optical Processor for Phased Array Antenna Beam Formation", Proc. of SPIE Vol 477 - 1984

2. Goodman Introduction to Fourier Optics McGraw Hill - Cpy 1968

3. Casaseut, et al "Phased Error Model for Simple Fourier Transform Lenses" Applied Optics Vol 17, No, 11, June 1978

4. Anderson, et al. "Antenna Beamforming using Optical Techniques - Follow on Study" NASA - LeRc Contract #NAS3-25720 Final Report - February 1991

5. Buckley, R. H. "Progress in Microwave Modulation Fiber Optic Links" Proc. SPIE Vol 1102, January 1979

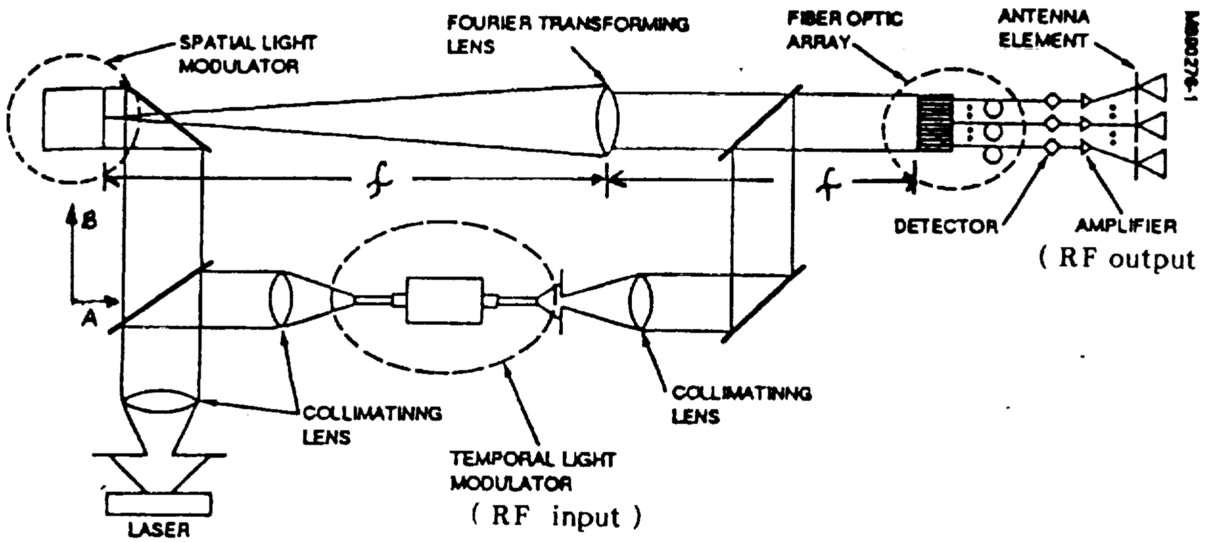


Figure 1 Optical Processor Beamfoming Network Schematic

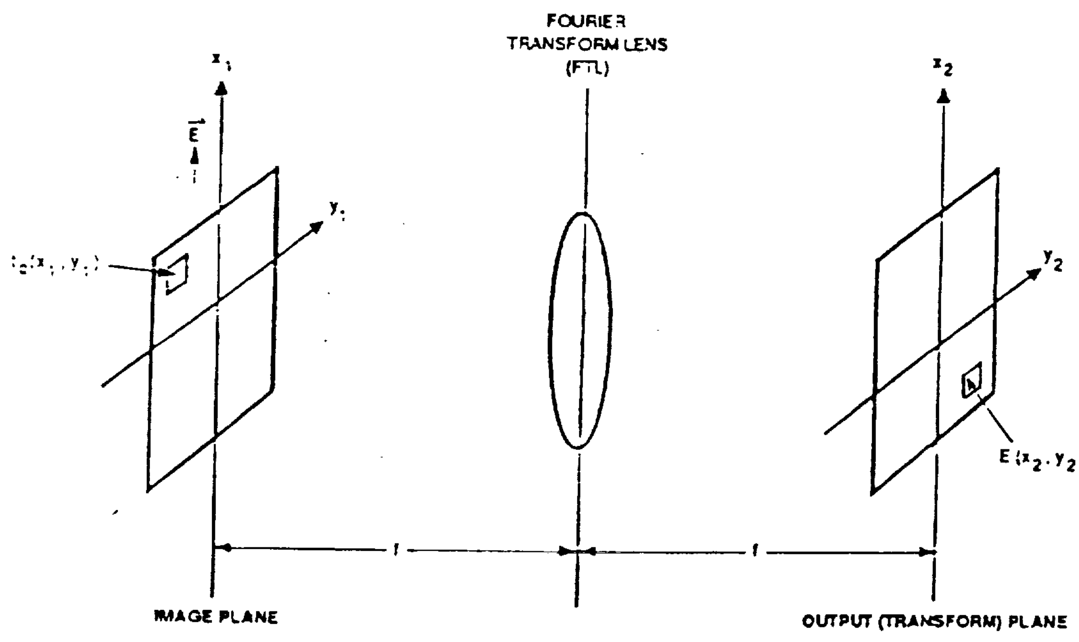


Figure 2 Spatial Processing Path Geometry

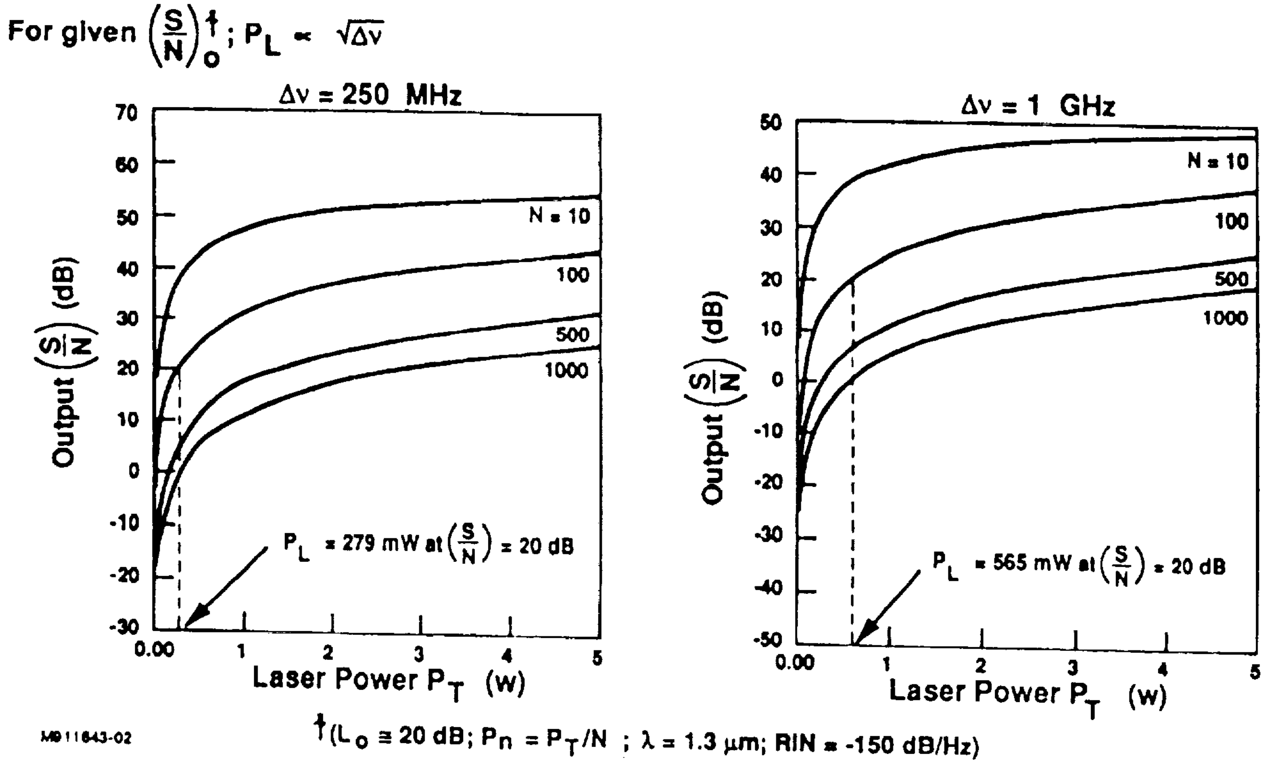


Figure 3 (S/N) Output vs. Required Laser Power

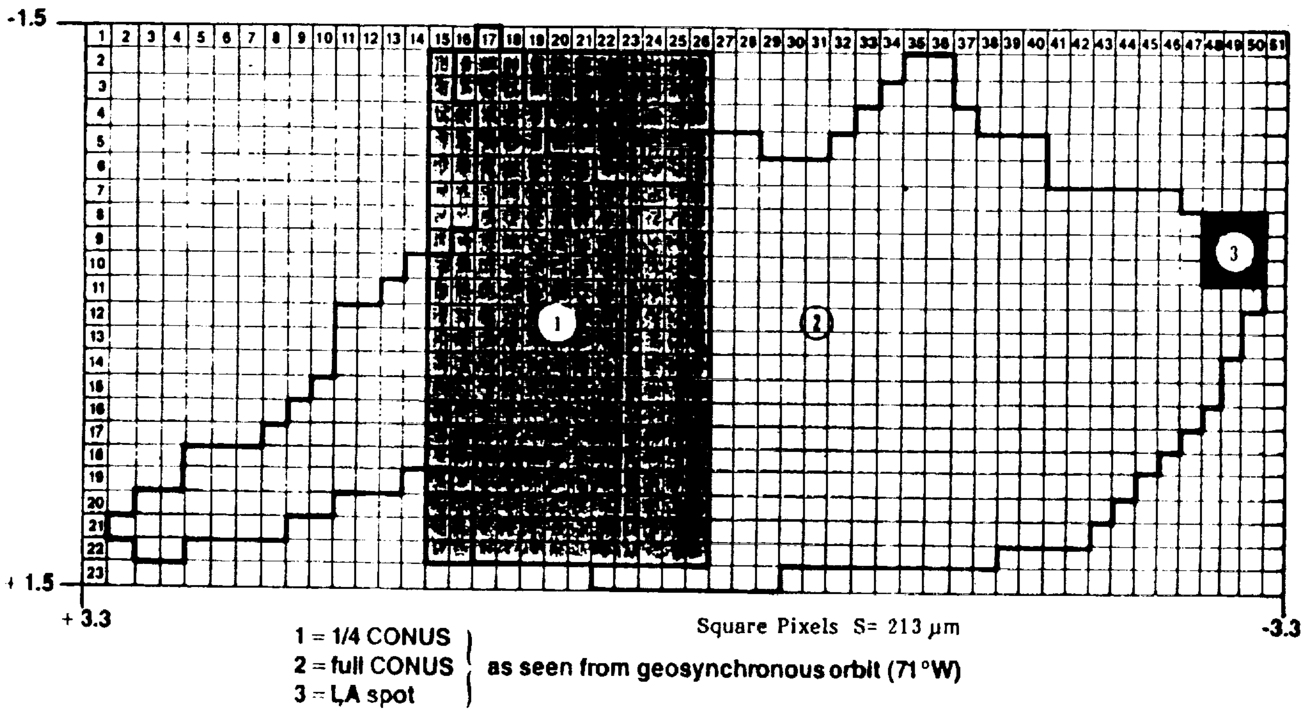


Figure 4 Liquid Crystal Light Valve (LCLV) Model

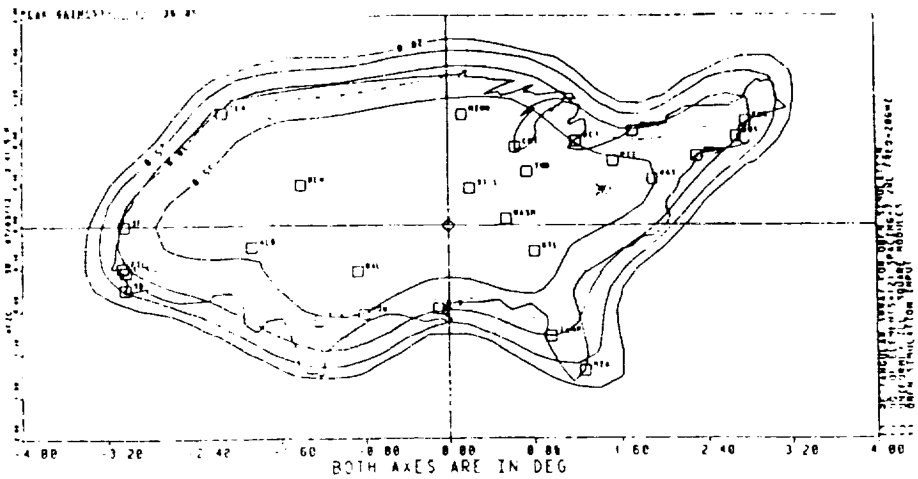


Figure 5a Full CONUS Beam Coverage

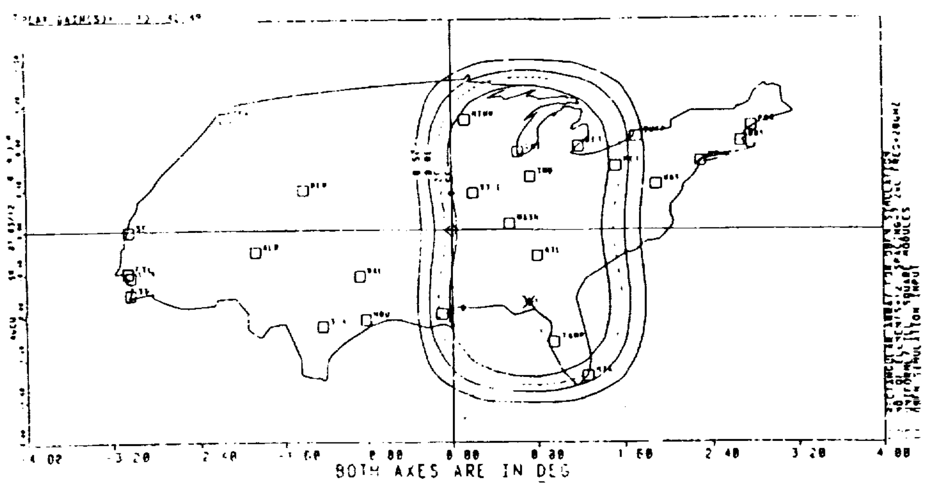


Figure 5b Midwestern Quarter CONUS Beam Coverage

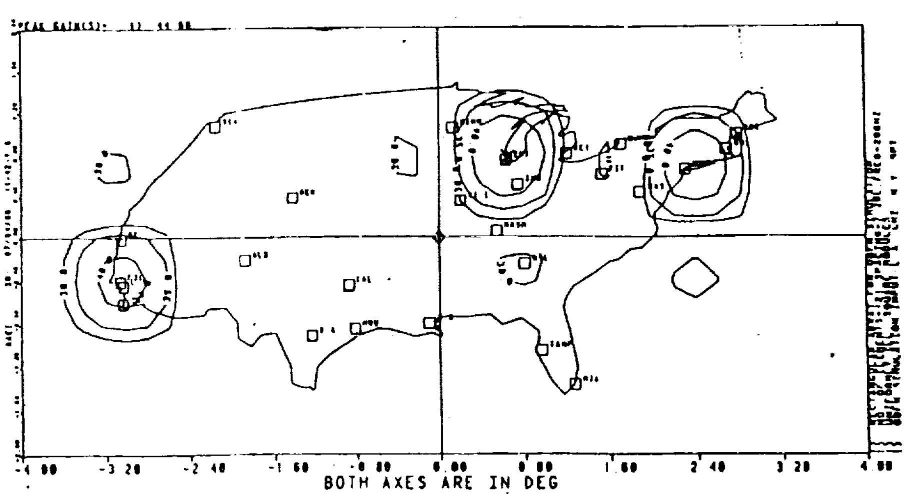


Figure 5c Multiple Spot Coverage (L.A./Chicago/N.Y.)

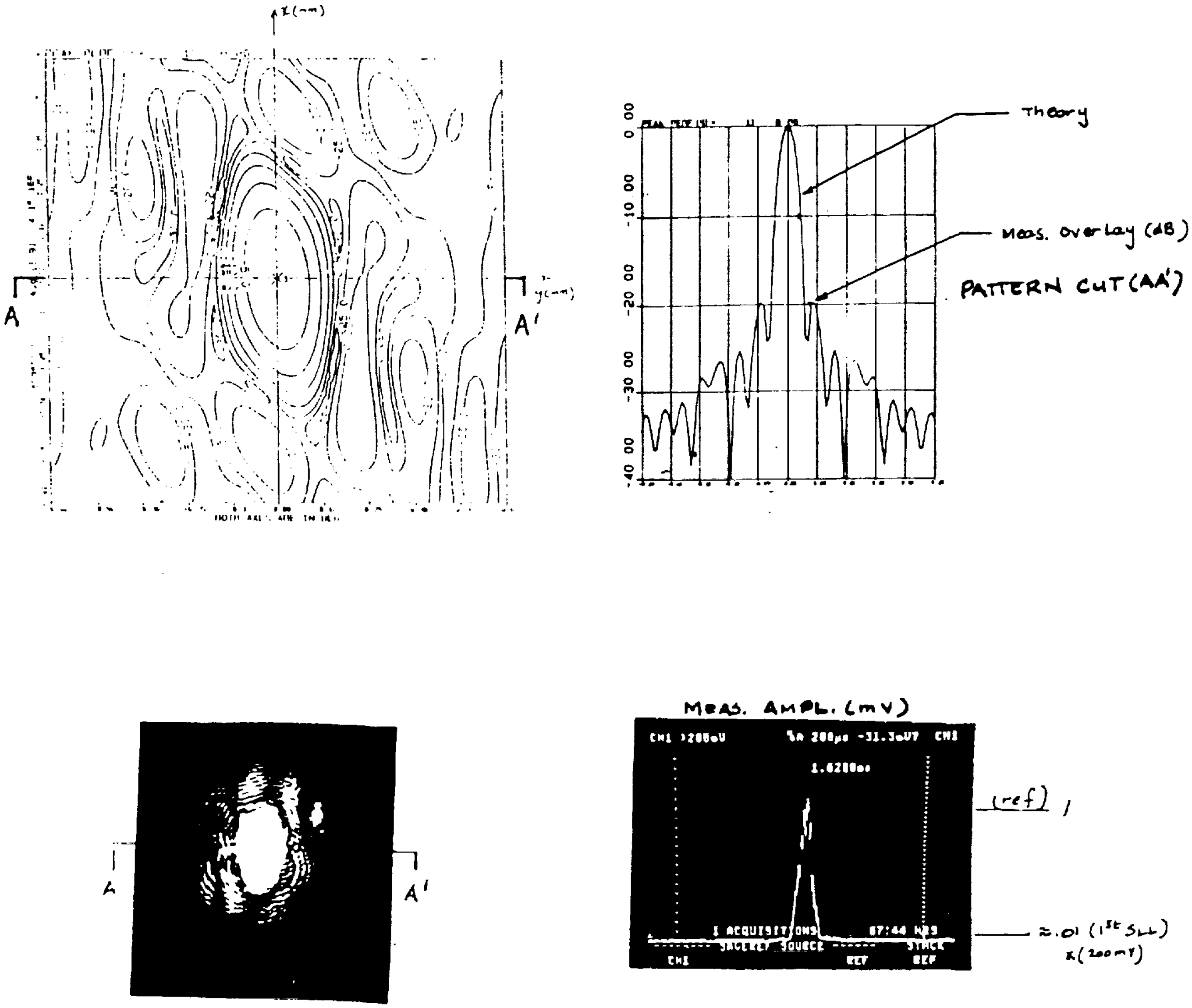
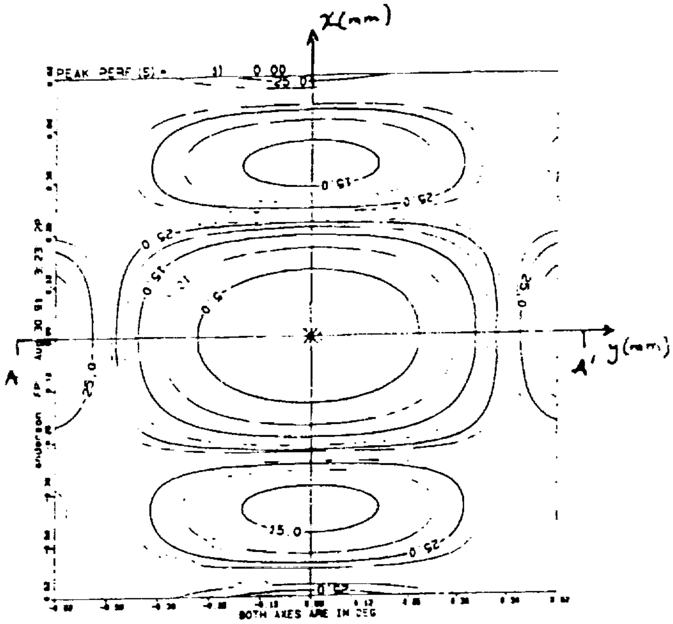
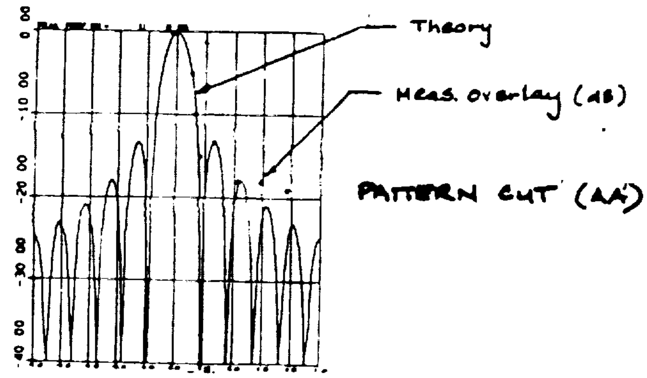
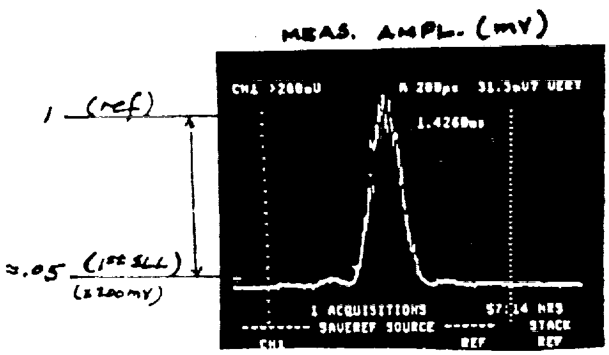


Figure 6 Computed vs. Measured CONUS aperture spectra using CCD camera and photographic film







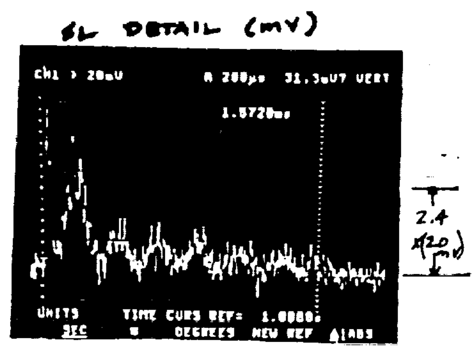


Figure 7 Computed vs. Measured Quarter CONUS aperture spectra using CCD camera with sidelobe detail

COMPUTER GRAPHICS AND SIMULATION

(Session A4/Room B1)

Tuesday December 3, 1991

- Global Positioning System Supported Pilot's Display
- Application of Technology Developed for Flight Simulation
- FAST: A Multi-Processed Environment for Visualization of Computational Fluid Dynamics
- A Full-Parallax Holographic Display for Remote Operations

N92-22436

GLOBAL POSITIONING SYSTEM SUPPORTED PILOT'S DISPLAY

Marshall M. Scott, Jr.
NASA John F. Kennedy Space Center
Kennedy Space Center, FL 32899

Temel Erdogan
Boeing Aerospace Operations
Kennedy Space Center, FL 32899

Andrew P. Schwalb
Boeing Aerospace Operations
Kennedy Space Center, FL 32899

Charles H. Curley
Boeing Aerospace Operations
Kennedy Space Center, FL 32899

ABSTRACT

This paper describes the hardware, software and operation of the Microwave Scanning Beam Landing System (MSBLS) Flight Inspection System Pilot's Display. The Pilot's Display is used in conjunction with flight inspection tests that certify the Microwave Scanning Beam Landing System used at Space Shuttle landing facilities throughout the world.

The Pilot's Display was developed for the pilot of test aircraft to set up and fly a given test flight path determined by the flight inspection test engineers. This display also aids the aircraft pilot when hazy or cloud cover conditions exist that limit the pilot's visibility of the Shuttle runway during the flight inspection. The aircraft position is calculated using the Global Positioning System and displayed in the cockpit on a graphical display. The runway, desired flight path, and "fly-to" needles are displayed for the pilot, as well as other information used by the flight inspection test engineers. A variation of the software and hardware also aids the pilot to fly inspection flights for Tactical Air Navigation ground stations at the Shuttle landing sites.

INTRODUCTION

Requirement

The Pilot's Display Subsystem is part of the Microwave Scanning Beam Landing System (MSBLS) Flight Inspection System and is used to coordinate the test requirements of the inspection team with the test aircraft pilot. The test requirements consist of radials (wherein the aircraft flies toward the runway at a constant altitude) and glide slopes (which are landing approaches at set angles; see figure 1). This display shows the pilot the correct flight path to fly

 PRECEDING PAGE BLANK NOT FILMED

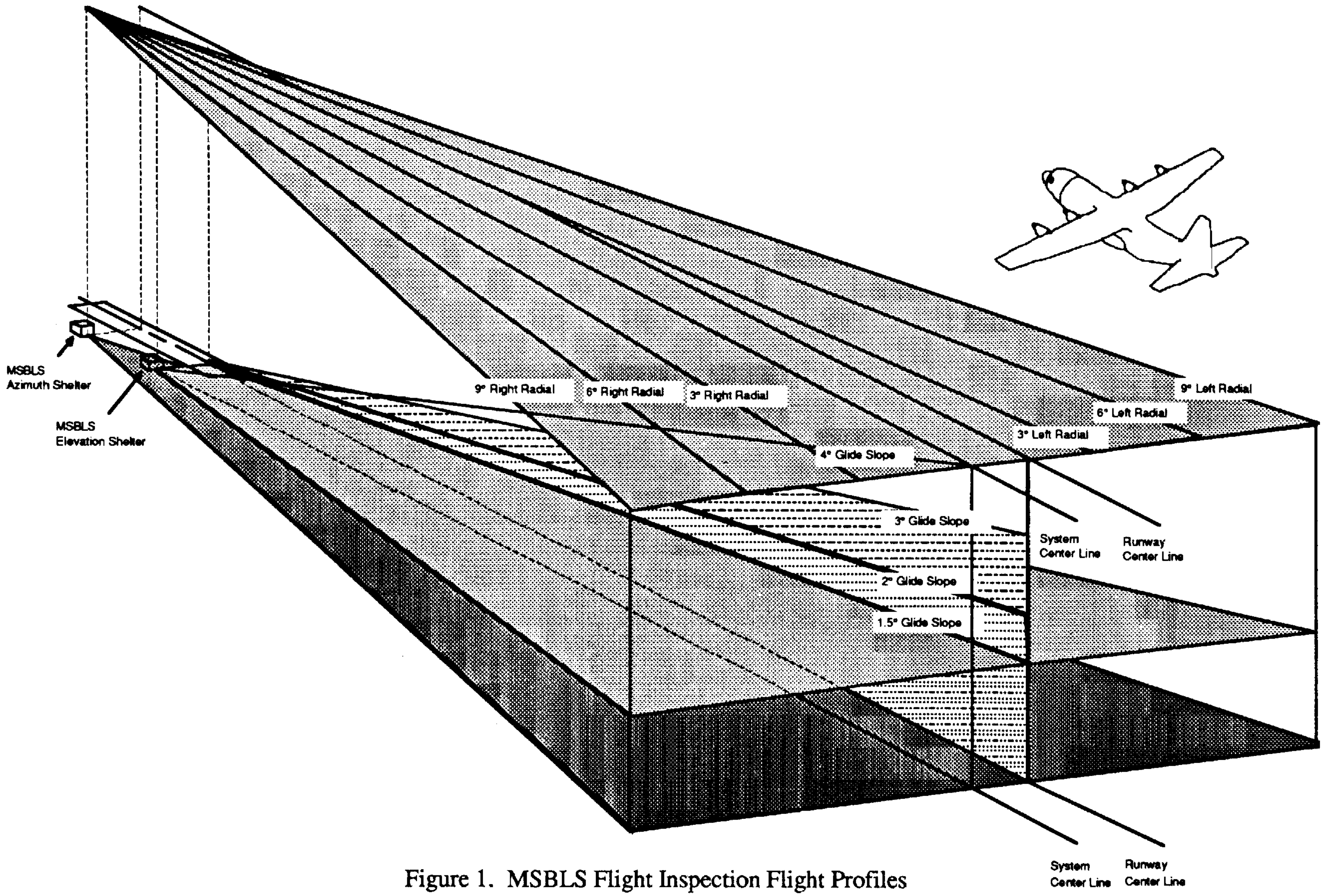


Figure 1. MSBLS Flight Inspection Flight Profiles

and provides "fly-to" needles so the aircraft can remain on the correct flight path during the different flight inspection test runs.

Aircraft Position Determination

The Pilot's Display, as well as the MSBLS Flight Inspection System, uses Global Positioning System (GPS) data to determine aircraft position in real-time. The GPS operates by the use of 21 satellites, when fully deployed, in 12-hour orbits that continuously broadcast their identification, position, and time in space. The position of the aircraft is determined by receiving the information transmitted by any four of the satellites and by computing the position using the orbital information the satellites provide and the time the information takes to travel from the satellite to the aircraft.

GPS Errors

The GPS satellites transmit information on two frequencies: 1575.42 MHz (L1) and 1227.60 MHz (L2). The L1 carrier contains a precision code (P-code) ranging signal and a coarse/acquisition (C/A) code. The L2 carrier contains only the P-code, which is intended for military use only. The GPS receivers used by the MSBLS Flight Inspection System can receive only the L1 frequency and process only the C/A code. The use of the two frequencies allows a GPS receiver that receives both L1 and L2 carriers to compensate for the effects of the ionospheric and tropospheric delays and greatly improve the accuracy of the computed position. This type of GPS receiver can also correct for another source of error called selective availability (SA). The SA is an intentional error placed in the satellite's information, and only receivers that can process the P-code are able to correct for this error in real-time.

Differential GPS

The effects of the GPS errors described in 1.3 and a few others can be reduced for those with only C/A receivers (receives only L1 carrier) by using a technique called Differential GPS, which is the technique used by the MSBLS Flight Inspection System. Differential GPS uses two C/A receivers, a receiver located in a user vehicle and a reference receiver at a known fixed location. Since both receivers see the same errors, the reference receiver calculates the errors from the information about its known location and transmit the error information to the user receiver to correct its calculated position. A set of GPS receiver equipment is located at the MSBLS elevation transmitter shelter (see figure 2) and a set is located in the aircraft with MSBLS receivers and decoders. This configuration of equipment sets up the GPS data in the differential mode for the MSBLS flight inspections.

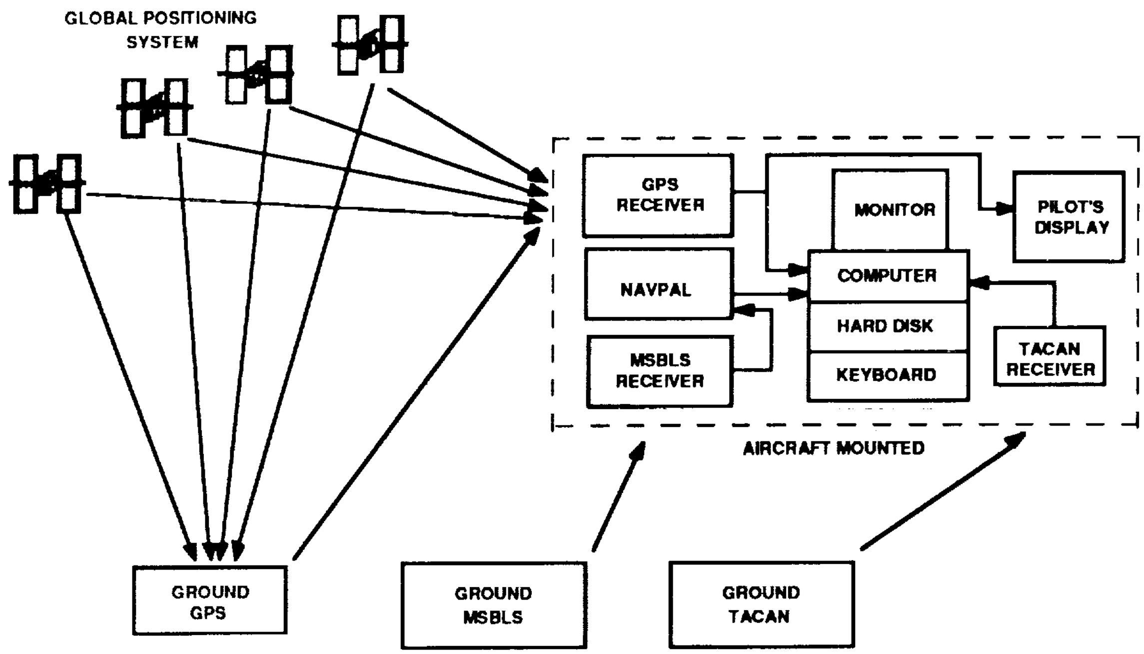


Figure 2. MSBLS Flight Inspection System

THEORY OF OPERATION

System Configuration

The Pilot's Display (see figure 3) receives the GPS position data from the same GPS receiver used by the MSBLS Flight Inspection System. The GPS receiver is initialized by the MSBLS Flight Inspection System computer and is set to receive and compute position data from satellites along with correction data from the reference receiver. The GPS receiver provides coordinate data (corrected X, Y, and Z in differential mode) to the system computer over an RS-232 serial data link at a data rate of 9,600 baud. The ground GPS receiver/modem (reference receiver) is initialized as the origin of the Pilot's Display coordinate system. The ground GPS receiver/modem will transmit position corrections for the coordinate system based on bias and drift errors in the GPS signal.

Display

The Pilot's Display computer is programmed to compute and display the aircraft's position in real-time on a video monitor. The display shows a pair of "fly-to" needles (see figure 4). The distance of each needle from the center of the display represents the aircraft's offset from the desired flight path. These offsets are computed once per second and are the differences of the aircraft's computed position and the chosen flight path. The pilot corrects the course of the aircraft by flying the aircraft so the needles move toward the center of the display as the error in the flight path is reduced.

Pilot's Monitor

The output of the computer color monitor card is sent to a graphics format conversion card (Folsom Research, Inc., Video 300 Card), which converts the computer graphics format to television NTSC format signals. The Video 300 card outputs a standard NTSC video signal which is sent over coax cable to the aircraft flight deck to the pilot's color monitor (5-inch JVC TM-63U).

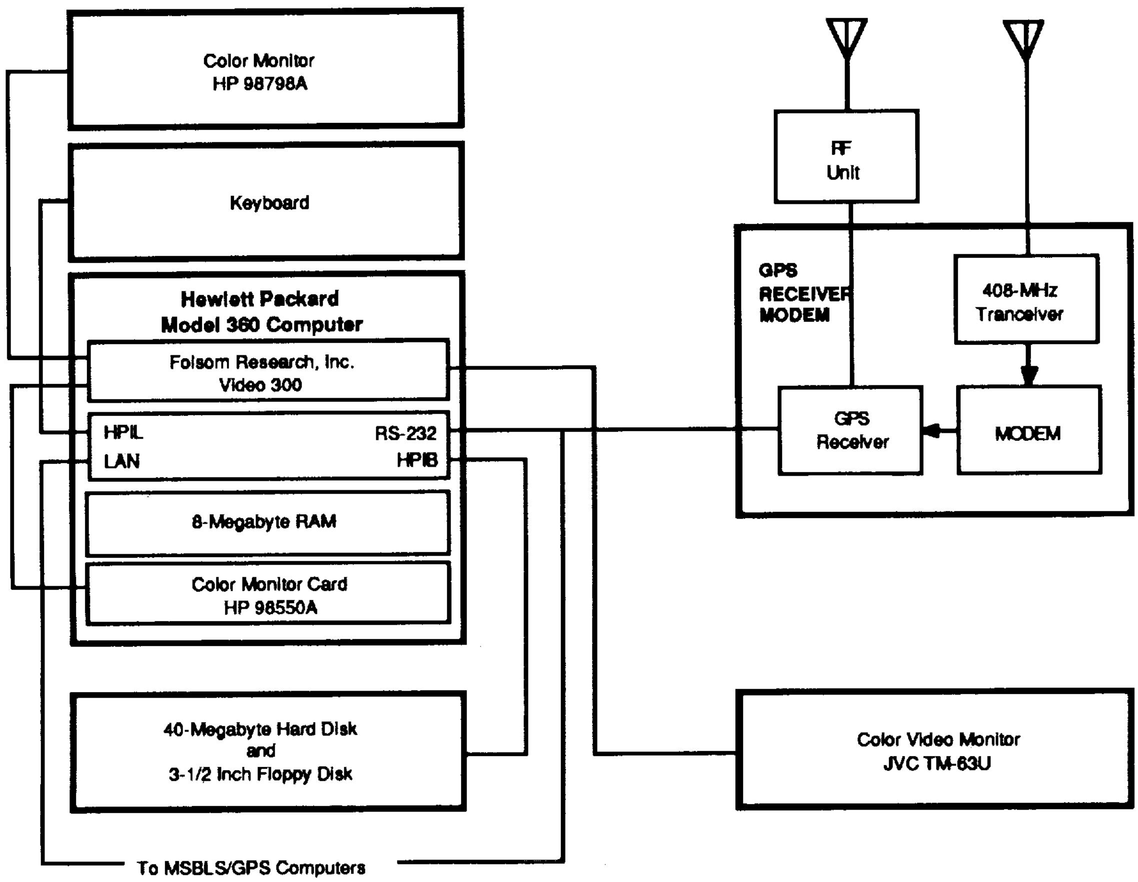


Figure 3. Pilot's Display System Block Diagram

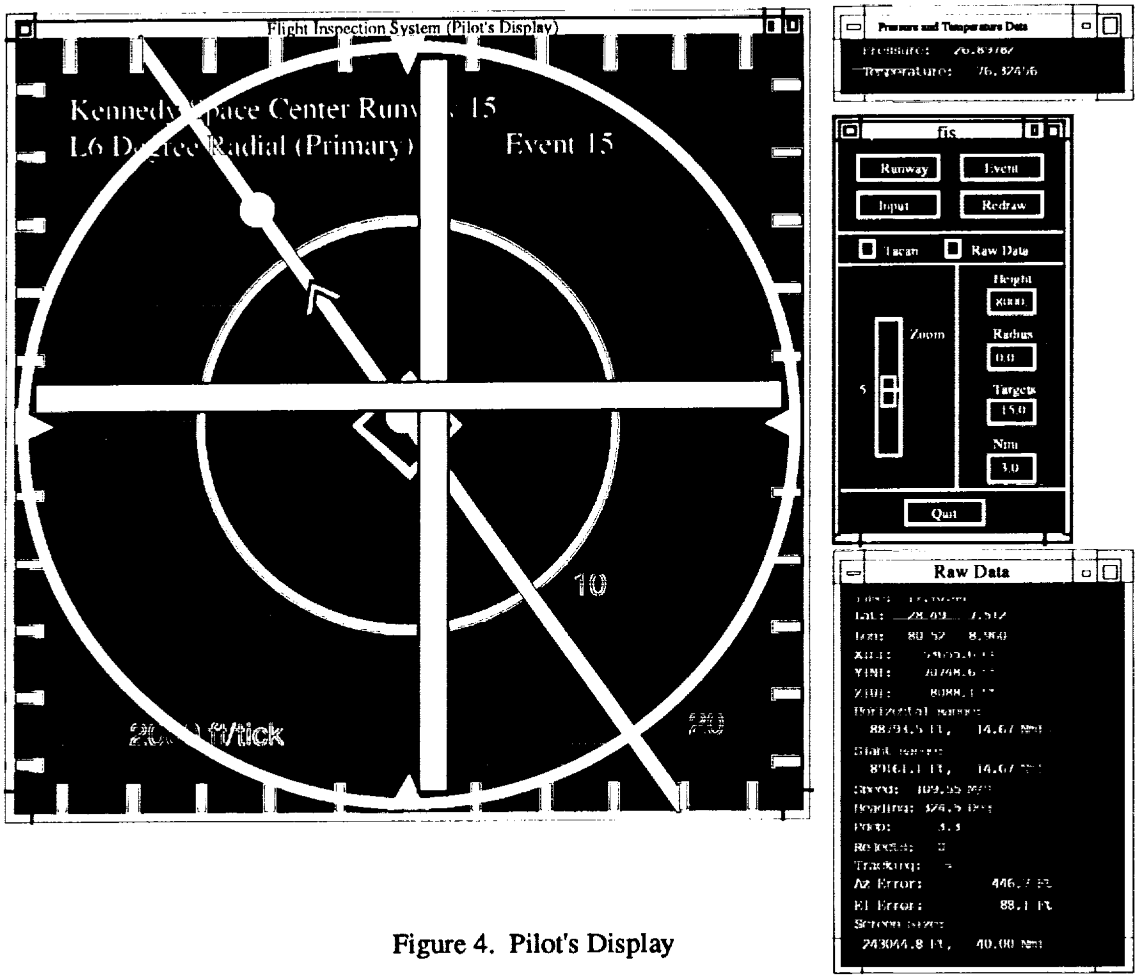


Figure 4. Pilot's Display

SOFTWARE DESCRIPTION

System Inputs

The pilot's display program uses GPS information from a GPS receiver to provide real-time updates of the position of the aircraft. To accomplish this, the pilot's display program reads the data coming from the GPS receiver via a 9600 baud serial link between the receiver and the pilot's display computer. The non-position information contained in the data stream is filtered out and the remaining information is used to determine the aircraft's position. This information appears about one a second in the data stream. The program can also replay GPS information previously recorded to a file. This file can only contain GPS position data and is not filtered in any way. The data will be replayed at about five times its normal rate of once a second.

The pilot's display computer is also connected to an analog to digital (A/D) converter that is connected to temperature and pressure transducers. The A/D converter is connected to the pilot's display computer via a HPIB link. During real-time data display it is queried whenever a position update is received on the GPS data stream (once a second) and the information is displayed to the operator and logged to a file, along with the current time (from the GPS data), for later analysis.

Operator Inputs

The pilot's display program runs in the X Windows Version 11 Release 3 (X11R3) environment and uses the Motif graphical user interface (GUI). When the program starts it presents the operator with an empty pilot's display and the operator's control panel. The program is controlled from the control panel via the mouse and/or the keyboard. The control panel has control buttons and sliders that control the operation of the pilot's display. A button is selected by either moving the mouse cursor on it and clicking the mouse button or by pressing the tab key on the keyboard until the desired button is highlighted and then pressing enter to select the button. The operator can perform the following program control actions from the control panel:

a) Select runway (Runway). Upon selection of this button, the operator will be presented a scrolling menu window of runways to select from. The runways are loaded from a database on the pilot's display computer and can be updated at any time to include new runways or to modify the characteristics of existing ones. Once a runway is selected, the pilot's display program will display the runway on the pilot's display and begin taking data from the currently selected GPS data stream, updating the aircraft's position as necessary. The runway database information was obtained from the Defense Mapping Agency Geodetic Survey Group, Detachment 4, Patrick Air Force Base, Florida, "Survey Results, Space Shuttle Support," dated January 1990.

b) Select event (Event). Upon selection of this button, the operator will be presented a scrolling menu window of flight path events to select from. The events are loaded from a database on the pilot's display computer and can be updated at any time to include new events or to modify the characteristics of existing ones. Each event is identified by an event number and description. If the currently selected mode is TACAN mode, then only TACAN events are displayed. If the mode is not TACAN, then all other events are shown. Once the event is selected, the flight path is displayed on the pilot's display, along with the "fly-to" needles.

c) Select input source (Input). This button allows the operator to select either real-time data mode or playback mode. This button will present a file selection window to the operator to allow the selection of another input source. If the operator selects the serial connection that the GPS real-time data is received then the program will operate in real-time mode. If the operator selects a file of recorded GPS data, the program will operate in playback mode. In real-time mode, GPS data is taken until the input source is changed. In playback mode, GPS data is taken from the file until the end of the file and no further updates occur until the operator selects another input source.

d) Redraw pilots display (Redraw). This button causes all the windows used by the pilot's display program to be cleared and redraw.

e) Select/Deselect TACAN mode (Tacan). This button controls whether TACAN events are selected when the Event button is pressed. This this button is in the pressed state TACAN mode is selected, else it is not.

f) Select to see raw GPS data (Raw Data). This button allows the operator to control whether a window displaying the raw GPS data is displayed. If the button is in the pressed state the a raw GPS data window will be displayed, else it will not be.

g) Change zoom factor (Zoom). This is a slider that has a range from one to ten and controls the scale of the pilot's display.

h) Change event characteristics (Radius, Height, Target In, Target Out). These text entry fields afford the operator the ability to change certain characteristics of the current event. These changes are only in effect as long as the current event remains selected.

Program Calculations

The program for the pilot's display calculates distances and errors by translating the raw GPS data and comparing it to the selected runway and event information. The raw GPS data gives position in longitude and latitude and this information is converted to local east, north and up (x,y,z) coordinates.

The program calculates two types of deviation: elevation and azimuth. Both are measured in feet. Elevation deviation is the distance that the aircraft is above or below the flight path. Azimuth deviation is the distance that the aircraft is to the left or right of the flight path.

Elevation deviations are calculated two different ways, each for a different type of flight event. If the event has no glide slope, then the elevation deviation is simply the difference between the desired height and the actual height. For glide slope events, the only difference is that the desired elevation has to be calculated based upon the aircraft's distance.

Azimuth deviations are handled in two different ways each for a different type of flight event. If the flight path is radial (i.e., a straight line), the difference between the aircraft's heading and the desired flight path heading is calculated rotating the aircraft's position about the system origin by an angle equal to the sum of the runway heading and the desired aircraft heading.

If the flight path is circular, azimuth deviations are calculated by determining the distance of the plane from the system origin and subtracting that from the desired distance. The sign of the deviation is determined by calculating the angular difference between the aircraft's heading and a tangent to the circular flight path.

CONCLUSIONS

MSBLS flight inspections have become more efficient with the Pilot's Display. The display coordinates the test engineer's requirements with the pilot of the test aircraft and provides the pilot a direct indication how well he is maintaining the test flight. This system has also been used to follow the progress of the aircraft while flying from one Shuttle Landing Facility to another, demonstrating the use of GPS as a navigation aid.

N92-22437

Application of Technology Developed for Flight Simulation at NASA Langley

Jeff I. Cleveland II

National Aeronautics and Space Administration
Langley Research Center
Hampton, Virginia 23665-5225

ABSTRACT

In order to meet the stringent time-critical requirements for real-time man-in-the-loop flight simulation, computer processing operations including mathematical model computation and data input/output to the simulators must be deterministic and be completed in as short a time as possible. In 1983, in response to increased demands for flight simulation performance, personnel from NASA's Langley Research Center (LaRC), working with KineticSystems Corporation, developed extensions to a standard input/output system to provide for high bandwidth, low latency data acquisition and distribution. The Computer Automated Measurement and Control technology (IEEE standard 595) was extended to meet the performance requirements for real-time simulation. This technology extension increased the effective bandwidth by a factor of ten and increased the performance of modules necessary for simulator communication. Under sponsorship of the LaRC Technology Utilization Office, the project team won an IR-100 award in 1986 for this system. This technology is being used by more than 80 leading technological developers in the United States, Canada, and Europe. Included among the commercial applications of this technology are nuclear process control, power grid analysis, process monitoring, real-time simulation, and chemical processing.

Personnel at LaRC are currently developing the use of supercomputers for simulation mathematical model computation for real-time flight simulation. This, coupled with the use of an open systems software architecture, will advance the state-of-the-art in real-time flight simulation.

This paper will describe the data acquisition technology innovation and the development of supercomputing for flight simulation.

INTRODUCTION

NASA's Langley Research Center (LaRC) has used real-time flight simulation to support aerodynamic, space, and hardware research for over forty years. In the mid-1960s LaRC pioneered the first practical, real-time, digital, flight simulation system with Control Data Corporation (CDC) 6600 computers. In 1976, the 6600 computers were replaced with CDC CYBER 175 computers. In 1987, the analog-based simulation input/output system was replaced with a high performance, fiber-optic-based, digital network. In 1990, action was begun to replace the simulation computers with supercomputers.

The digital data distribution and signal conversion system, referred to as the Advanced Real-Time Simulation System (ARTSS) is a state-of-the-art, high-speed, fiber-optic-based, ring network system. This system, using the Computer Automated Measurement and Control (CAMAC) technology, replaced two twenty year old analog-based systems. The ARTSS is described in detail in references 1 through 6.

An unpublished survey of flight simulation users at LaRC conducted in 1987 projected that computing power requirements would increase by a factor of eight over the coming five-years (Figure 1). Although general growth was indicated, the pacing discipline was the design testing of high performance fighter aircraft. Factors influencing growth included: 1) active control of increased flexibility, 2) less static stability requiring more complex automatic attitude control and augmentation, 3) more complex avionics, 4) more sophisticated weapons systems, and 5) the need to simulate multiple aircraft interaction, the so called "n on m" problems. This requirement for more computing power is at least, if not industry wide, common to the fighter aircraft segment.

Two single processor Control Data Corporation CYBER 175 computers, tightly coupled through extended memory, are used to support flight simulation. Having decided to continue using large-scale general-purpose digital computers, LaRC issued a Request for Proposals in May, 1989 and subsequently awarded a contract to Convex Computer Corporation in December of that year. The resulting computational facility provided by this contract is the Flight Simulation Computing System (FSCS).

ADVANCED REAL-TIME SIMULATION SYSTEM

Through design efforts by both LaRC design engineers and design engineers at KineticSystems Corporation, three components of the ARTSS were developed to meet LaRC requirements. These were the serial highway network, the network configuration switch, and the signal conversion equipment. A block diagram of the ARTSS is presented in Figure 2.

Serial Highway Network

The LaRC ARTSS employs high-speed digital ring networks called CAMAC highways. At any given time, four totally independent simulations can be accommodated simultaneously. An aircraft model is solved on one of the mainframe computers and the simulation is normally assigned one highway. The purpose of the network is to communicate data between the central computers and simulation sites (control console, cockpit, display generator, etc.). The elements of a CAMAC highway are: the Block Transfer Serial Highway Driver (BTSHD); the fiber-optic U-port adaptor, the Block Transfer Serial Crate Controller (BTSCC); the List Sequencer Module (LSM); and the CAMAC crate. Three features of the networks were developed to meet the LaRC requirement. First, the mainframe computer interface to the BTSHD was developed. Second, the block transfer capability was developed to meet LaRC performance requirements. This capability resides in the BTSHD, BTSCC, and LSM. Third, the fiber optic capability was developed to satisfy our site distance problem. The simulator sites are from 350 to 6,000 feet from the computer center.

Prior to the development of the block transfer capability, a CAMAC message was approximately 19 bytes long which included addressing, 24 bits of data, parity information, and response information. The addition of the block transfer capability allowed for the inclusion of many CAMAC data words in a single message. During block transfers, data reads or writes proceed synchronously at one 24-bit CAMAC data word per microsecond. This is several times faster than the normal single word message rate. Besides the CAMAC standard message, there are two modes of block transfer. In the first, the entire block of data goes to a single module within a crate. It is implemented by the BTSCC repeating the module-select and function bits on the crate dataway for each CAMAC word. In the second block transfer mode, the block, either on read or write, is divided among several modules within a crate. This mode employs the LSM module which is loaded by the mainframe computer at set-up time with up to four lists of module-select and function bits. When this type of block transfer is in progress, the BTSCC acquires module number, function, and subaddress for each sequential CAMAC word in the block from the indicated list in the LSM.

Network Configuration Switch

The purpose of the network configuration switch system is to provide complete connectability between the simulation applications on the mainframe computer and the various simulation sites. Upon request, any sensible combination of available sites can be combined into a local CAMAC ring network in support of a single simulation. This network configuration for a given simulation is done during the initialization phase, after a highway has been assigned by the scheduling software. The application job requests sites by resource request statements and if the sites are available, the switch will electrically and logically configure the network without disturbing other running simulations. The switch is built for a maximum of 12 highways and 44 sites. Each highway may be connected to a different host computer.

Signal Conversion Equipment

Three types of output converter modules and two types of input modules were designed and built to LaRC specifications. The converters are high quality and have been added to the vendor's catalog. The digital-to-analog converters (DAC), analog-to-digital converters (ADC), and digital-to-synchro (DSC) are 16-bit devices with 14 bits of accuracy. The data transmitted uses 16 bits although only 14 bits are meaningful. This implementation allows LaRC to change converter precision without major changes in software or protocol. To decrease transmission time, data words are packed such that three converter words (16 bits each) are contained in two CAMAC words (24 bits each). The discrete input converters contain 48 bits per module and the discrete output modules contain 24 bits per module.

Clocking System

Flight simulation at LaRC is implemented as a sampled data system. The equations of motion are solved on a frame-by-frame basis using a fixed time interval. To provide the frame interval timing signals and a clocking system for synchronization of independent programs, LaRC designed and built the real-time clocking system. This system is patented and is described in reference 7. The clock system is composed of a central unit and multiple CAMAC modules called Site Clock Interface Units (SCIUs) which are connected by means of a separate fiber optic star network. Two distinct time intervals are broadcast by the central unit on a single fiber. The first time interval has a constant 500-microsecond period. The tic count necessary for a real-time frame is set in the SCIU by initialization software. This count is decremented by one for each occurrence of the interval timer. When the count reaches zero, each SCIU issues a signal that indicates beginning of frame. The frame time is determined independently for each simulation but must be a multiple of 500 microseconds. The second clock signal, called the job sync tic, has a longer period called the clock common multiple which is set manually. This longer period is used for synchronization. Each frame time must divide evenly into the clock common multiple, ensuring that all simulations will be synchronized on the occurrence of the job sync tic.

REQUIREMENTS FOR NEW SIMULATION COMPUTING SYSTEM

In 1987 LaRC conducted a survey of simulation users and program managers to determine future requirements for flight simulation at LaRC. Results from this study (see Figure 1) indicated that high performance aircraft with expected increasingly complex models and flexible airframes would require up to eight times the model computing capacity compared to the present two CYBER 175 computers. These results coupled with other information, led to the definition of requirements for replacing the existing computers. Following an extensive survey of the marketplace, further action was delayed for one year to wait for the development of systems that could meet the requirements. In 1989 the requirements were incorporated into a Statement of Work and a formal Request for Proposals was issued. The resulting system is called the Flight Simulation Computing System (FSCS).

CPU Performance

Real-time flight simulation at LaRC requires high scalar CPU performance to solve the equations of motion of the system being simulated. Using an existing X-29 simulation as a benchmark, the following CPU performance was specified:

1. If a single CPU configuration is provided, the CPU must solve the benchmark in at most 165 seconds.
2. If a multiple CPU configuration is provided, each CPU must solve the benchmark in at most 330 seconds.

Due to secure processing requirements, a minimum of two and a maximum of four independent computers were required. The CYBER 175 computer solves the benchmark in 660 seconds. Thus, the capabilities of the resultant total system will provide at least eight times the CPU processing power of the coupled CYBER 175 computers.

Real-Time Input/Output

The ARTSS CAMAC system has provided LaRC with a high performance real-time input/output system that has extended the capabilities of the LaRC simulation system. Since ARTSS provides a high transfer rate with low latency, LaRC required provision of a compatible interface between the simulation computing system and the ARTSS CAMAC system. LaRC required that the new system include all software and hardware to connect to the ARTSS CAMAC real-time network. This connection was required to transfer block data over the network at a sustained rate of 24 million bits per second in the enhanced serial mode.

Responsiveness

One of the critical requirements for any real-time simulation system is system responsiveness. The FSCS system is required to respond to an external event, cause a short FORTRAN program to execute, and post an observable output response, in less than 150 microseconds. This elapsed time, called time-critical system response, is measured at an external port on the computer. The external event occurs at a repetitive rate of 1000 events per second. In addition to the time-critical system response, CAMAC input/output response is required to be less than 200 microseconds. CAMAC input/output response is defined as the time between the action of an interrupt generated in a CAMAC crate, transfer of one CAMAC word of data, execution of the short FORTRAN program, and transfer of one CAMAC word of output.

Frame Rate

To support simulation applications needing higher frame rates, LaRC required the system to run simulations at 1000 frames per second. At this frame rate, during any given frame, the system must deliver at least 600 microseconds of CPU time for the simulation model with 100 bytes of real-time input and 100 bytes of real-time output. The sum of system overhead and real-time input/output must be less than 400 microseconds.

Real-Time Data Recording and Retrieval

To support real-time data recording and retrieval during synchronous flight simulation, LaRC required the capability to record and/or retrieve information from two files for each simulation. The aggregate storage capacity was required to be a minimum of 180 megabytes. Sufficient data rate was required to permit a simulation to record or retrieve one 1000-byte record per real-time frame from each file simultaneously at a frame rate of 100 frames per second.

Language and other factors

At LaRC, almost all simulation programs are written in the FORTRAN language. Furthermore, simulations have been developed on CDC 6000 series computers and succeeding generations for over twenty years. With simulations written taking advantage of the CDC 60-bit architecture, LaRC required that the FSCS system support simulations written in the FORTRAN language with a minimum floating point mantissa precision of 14 decimal digits and with a minimum exponent range of plus and minus 250 decimal. In addition, the C language is required to support a limited number of applications, and Pascal is required to support the CAMAC configuration database.

An application development capability was required to operate simultaneously with simulations operating in real-time using all the real-time computing power specified. This application development capability has a minimum performance specification and required an advanced source language level debugger.

NEW SIMULATION COMPUTING SYSTEM

The computers that LaRC is putting in place to fulfill the requirements are Convex Computer Corporation C3200 and C3800 series computers. These computers are classified as supercomputers and support both 64- and 32-bit scalar, vector, and parallel processing technology. The new system will be delivered in stages as the software system evolves. The first delivery consisted of a Convex C3230 (3 CPUs expandable to 4) with two CAMAC interfaces. The system was delivered with two peripheral buses (PBUS): one PBUS that is used for input/output to standard peripherals such as tape, disk, and line printer and one PBUS that is used exclusively for real-time input/output to the ARTSS CAMAC network. Each VME Input/Output Processor (VIOP) is a Motorola 68020 microcomputer that provides programmable input/output control. Each VIOP is connected to a standard 9U VMEbus and to the corresponding PBUS. The CAMAC interface consists of a KineticSystems Model 2140 Enhanced Serial Highway Driver for VMEbus. The second delivery will consist of one Convex C3840 (4 CPUs expandable to 8) computer configured similar to the C3230 with 2 PBUSs and two CAMAC interfaces. The computer will contain 512 megabytes of main memory and sufficient disk and other peripherals to support flight simulation. The resulting computer configuration will be as shown in Figure 3.

There are four critical aspects of a computing system to support real-time simulation. These are: CPU performance, memory capacity, time-critical system response, and deterministic system performance.

The first computer installed (C3230) performs a simulation of an X-29 aircraft in 245 seconds per CPU which is 2.7 times faster than the computers being replaced. With two CPUs available for real-time, this results in 5.4 times the CPU performance. The second computer (C3840) performs the X-29 in 117 seconds per CPU which is 5.6 times faster than the computers being replaced. With three CPUs available for real-time, this results in 17 times the CPU performance.

Memory capacity is more than adequate to meet the requirements. The expanded memory capacity, compared with the old system, has allowed LaRC researchers to greatly increase the complexity of the simulations. The increase in memory capacity, coupled with the increase in CPU performance, has led to much higher fidelity simulations.

Time-critical system response is a measure of how fast the computing system can respond to real-time events from outside the computing system. Time-critical system response on both the computing systems has been measured at 31 microseconds which exceeds the LaRC requirement.

Deterministic system performance is a measure of how consistently on a frame-by-frame basis the computing system calculates the simulation model without any loss in synchronization with real-time. To use a computing system for real-time simulation, the system must be able to solve the model in a very nearly fixed amount of time, no matter what the demands on the system are for other computing. Both the computing systems have been measured to be deterministic within two percent which is excellent.

Operating System

Convex Computer Corporation offers two real-time operating systems. The operating system currently in use at LaRC requires one CPU for all non-real-time activity: editing, program compilation, and other UNIX activities. The other CPUs may be dedicated to real-time simulation. At the request of a real-time program, the program is locked down in memory to prevent page faults and the CPU or CPUS are dedicated exclusively to the real-time program.

The second real-time operating system incorporates a specially developed real-time kernel that the entire operating system is built upon. With this version of the real-time operating system, the UNIX operating system portion will be pre-empted by real-time requests and the response to real-time interrupts will be deterministic and very short. This version supports, on a single CPU, all activities of a normal UNIX operating system and also simultaneously supports real-time applications. This operating system requires special hardware that is not available in LaRC computers.

CONCLUSION

NASA Langley Research Center is at the mid-point in the development of a system to simulate in real-time increasingly complex and high performance modern aircraft. Utilizing centralized supercomputers coupled with a proven real-time network technology, scientists and engineers are performing advanced research using flight simulation. Hardware and software developed and concepts used are applicable to a wide range of commercial applications that require time-critical computer processing including process control, data acquisition, and real-time simulation of a wide variety of systems.

REFERENCES

1. Crawford, D. J. and Cleveland, J. I. II, "The New Langley Research Center Advanced Real-Time Simulation (ARTS) System," AIAA Paper 86-2680, October 1986.

2. Crawford, D. J. and Cleveland, J. I. II, "The Langley Advanced Real-Time Simulation (ARTS) System," AIAA Journal of Aircraft, Vol 25, No. 2, February 1988, pp. 170-177.

3. Crawford, D. J., Cleveland, J. I. II, and Staib, R. O., "The Langley Advanced Real-Time Simulation (ARTS) System Status Report," AIAA Paper 88-4595-CP, September 1988.

4. Cleveland, J. I. II, Sudik, S. J., and Crawford, D. J., "High Performance Processors for Real-Time Flight Simulation," AIAA Paper 90-3140-CP, September 1990.

5. Cleary, R. T., "Enhanced CAMAC Serial Highway System," presented at the IEEE Nuclear Science Symposium, San Francisco, California, October 23-25, 1985.

6. ANSI/IEEE Standards 583, 595, and 675, Institute of Electrical and Electronic Engineers, 1976.

7. Bennington, D. R., "Real-Time Simulation Clock," LAR-13615, NASA Tech Briefs, June 1987.

Flight Simulation Research Requirements

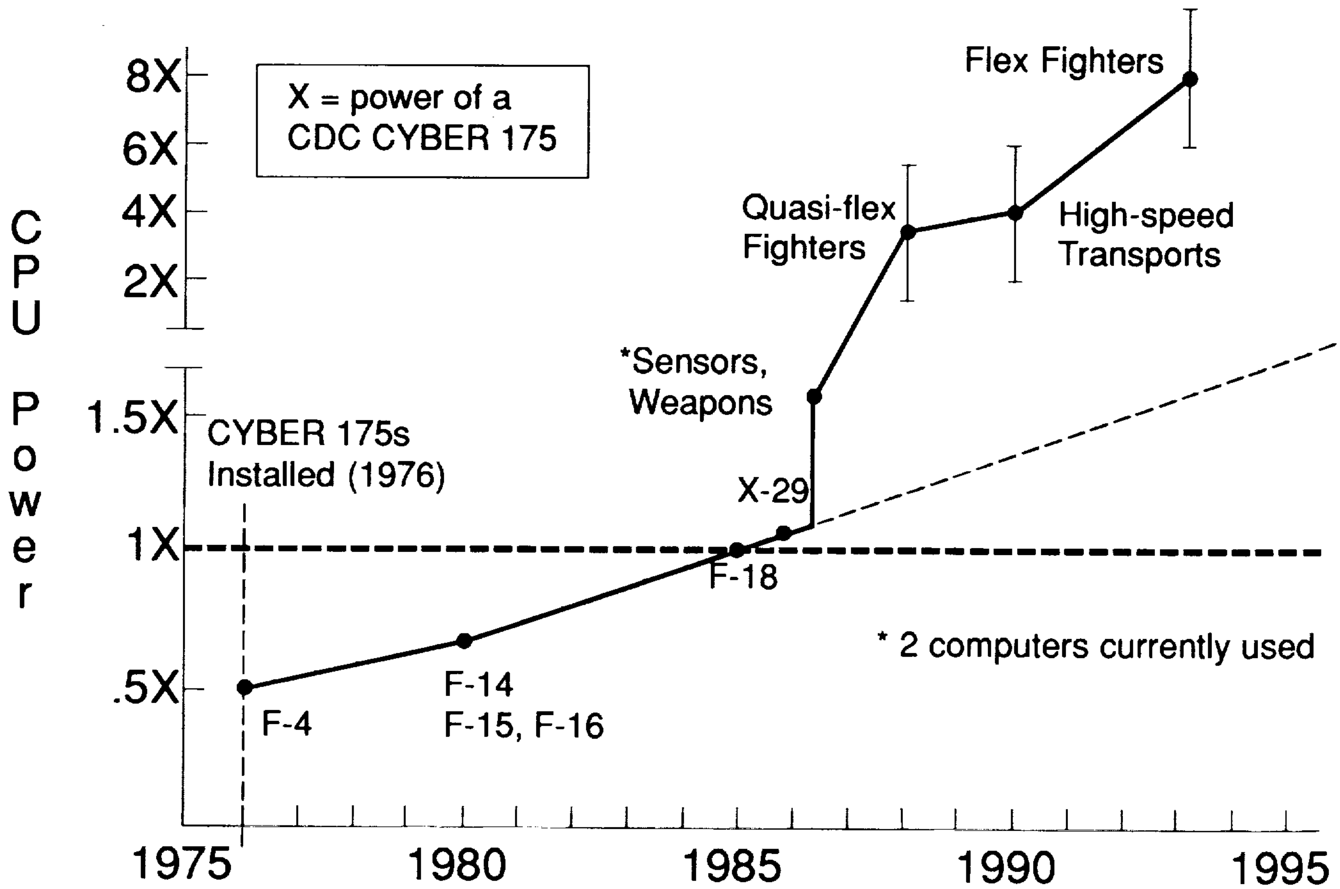
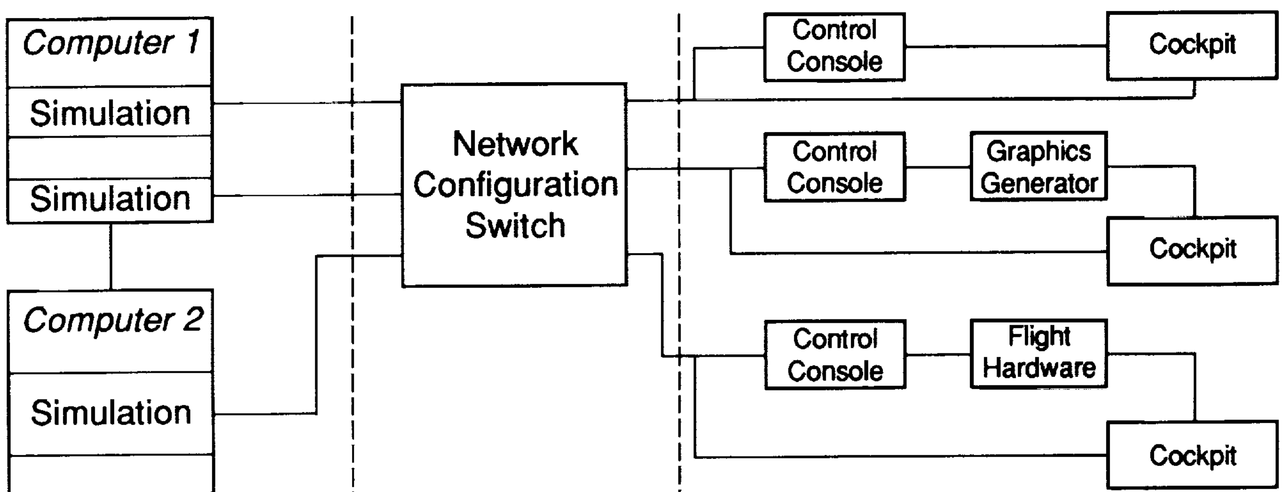


Figure 1.

Flight Simulation Subsystem



- Two CDC Cyber 175's
- Integrating two Convex computers with Unix
- Up to four simultaneous simulations
- ANSI standard FORTRAN simulation applications

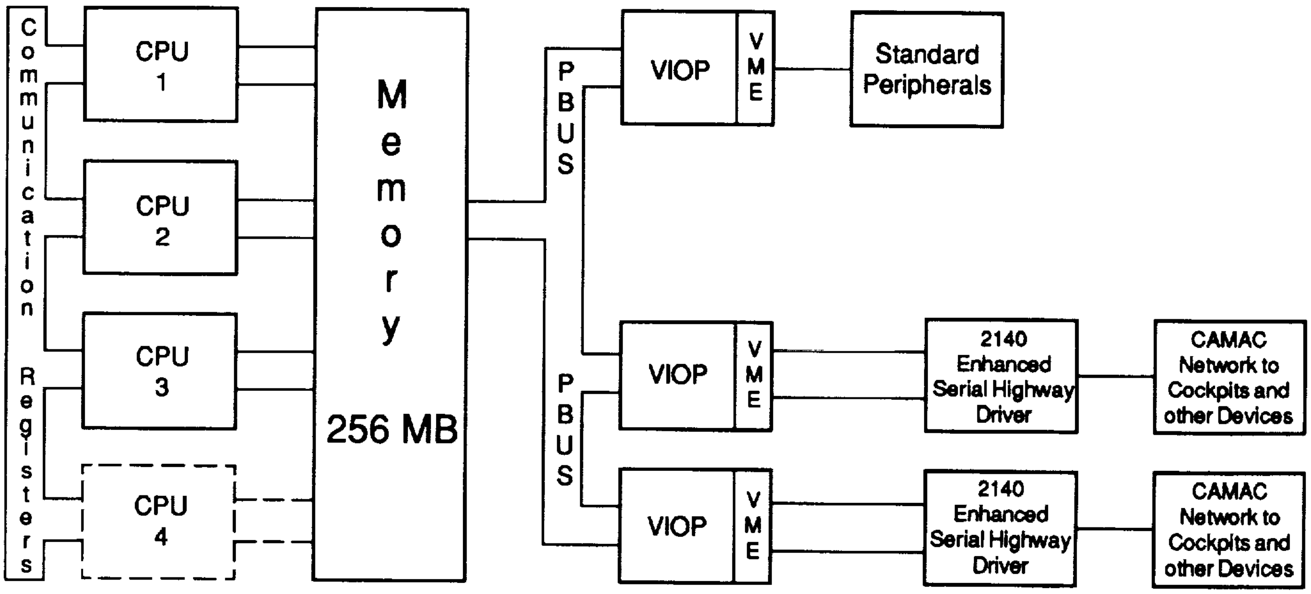
- Up to 12 networks
- Up to 7 computers
- Up to 44 simulator sites
- Computer controlled dynamic reconfiguration

- 50 mbit fiber optic network
- Up to 6000 feet between elements
- Up to 62 CAMAC crates per network
- Up to 25 addressable modules per crate
- No buffering in network yielding very low latency

Figure 2.

Computing System Configuration

Convex C3230 Computing System



Convex C3840 Computing System

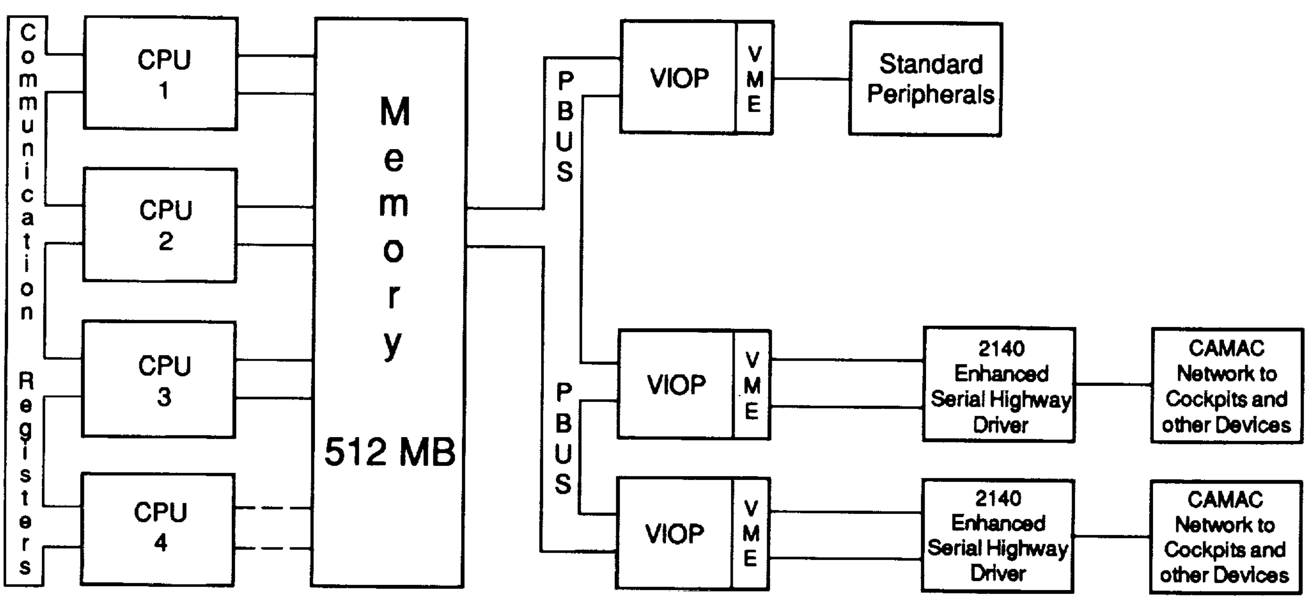


Figure 3.

N92-22438

FAST[1]: A MULTI-PROCESSED ENVIRONMENT FOR VISUALIZATION OF COMPUTATIONAL FLUID DYNAMICS

Gordon V. Bancroft
Fergus J. Merritt
Todd C. Plessel
Paul G. Kelaita
R. Kevin McCabe

Sterling Federal Systems Inc.
1121 San Antonio Road
Palo Alto, California 94303

ABSTRACT

Three-dimensional, unsteady, multi-zoned fluid dynamics simulations over full scale aircraft is typical of problems being computed at NASA Ames' Numerical Aerodynamic Simulation (NAS) facility on CRAY2 and CRAY-YMP supercomputers. With multiple processor workstations available in the 10-30 Mflop range, we feel that these new developments in scientific computing warrant a new approach to the design and implementation of analysis tools. These larger, more complex problems create a need for new visualization techniques not possible with the existing software or systems available as of this writing and these visualization techniques will change as the supercomputing environment, and hence the scientific methods employed,[1] evolve even further.

Visualization of computational aerodynamics requires *flexible, extensible*, and *adaptable* software tools for performing analysis tasks. *Flexible* means the ability to handle a diverse range of problems. *Extensible* means the ability to interact at all levels of the software hierarchy, either through existing built-in functionality or through the implementation of custom "plug-in" modules. *Adaptable* means the ability to adapt to new software and hardware configurations through the use of modular structured programming methods, a graphics library standard, and the use of common network communication protocols (like UNIX sockets) for the distribution of processing.

This paper discusses FAST (Flow Analysis Software Toolkit), an implementation of a software system for fluid mechanics analysis that is based on this approach.

BACKGROUND

Computational Fluid Dynamics (CFD), involves the use of high speed computers to simulate the characteristics of flow physics. Computational aerodynamicists use CFD methods and solvers to study subsonic, supersonic, transonic and hypersonic (compressible) regimes of flight, in addition to studying incompressible problems within particular systems. Examples of ongoing studies on full-scale aircraft configurations at NASA Ames include the Space Shuttle, F16, and the Aerospace Plane. Specialized areas of research include jet-engine turbine flow, VSTOL and ground effect research, and even flow through an artificial heart. Basic CFD research involves unsteady flow phenomena like vortex shedding and turbulence modelling.

A flow solver running on a supercomputer must handle input files (finite difference grids, ref. 7,15,16) that are typically very large. For example, the number of xyz triplets (each represented by three eight-byte floating point numbers) in a 100 x 100 x 100 grid yields a 24 Mbyte file. If complexity is added, or the grid resolution (density of points) must be raised for flow solving to yield acceptable results, the files grow proportionally in size. Once the solver has been run, there are from five to eight variables for each grid node, again, each represented by an eight-byte floating point number. For the 24 Mbyte example, five variables for each grid point yields a 40 Mbyte raw data file. This is a total of 64 Mbytes (grid plus the solution) for this example. The F16 mentioned previously, which consists of 29 grid zones, is over 108 Mbytes worth of data! *(Note: On the workstations these become four-byte IEEE format floating point numbers making the files about half this size)*

---

1 FAST (Flow Analysis Software Toolkit) Developed by Sterling Federal Systems Inc. under contract to NASA Ames Research Center NASA Contract #NAS2-13210.

Three examples of grid generation programs are:

3DGRAPE	3-dimensional grids about anything by Poissons Equation (Sorensen)
EAGLE	
GRIDGEN3D	

A list of commonly used flow solvers are:

ARC2D	Ames Research Center 2-dimensional solver (Pulliam)
ARC3D	Ames Research Center 3-dimensional solver (Pulliam)
TNS	Transonic Navier Stokes solver (Flores)
CNS	Compressible Navier Stokes solver (Flores)
PNS	Parabolized Navier Stokes solver (Chausee)
INS3D	Incompressible Navier Stokes solver (Kwak)
TWING	Transonic Wing solver (Thomas)

Programs available for visualization of CFD data sets are:

PLOT3D	A command line driven Fortran program that computes CFD quantities (Buning [7])
SURF	Allows for the rendering of smooth, wireframe, and function mapped surfaces with a more interactive interface (Plessel[8])
GAS	Combines graphics generated from PLOT3D and SURF and allows animations to be created and recorded (Merritt[9])
RIP	A program for interactive particle tracing (Rogers[19])

FAST OVERVIEW

The software cycle for the creation and and analysis of computational fluids results could be reduced to the following conceptual model:

- Data generation *(Flow solving)*
- Data manipulation *(The original data may need to be filtered or transfered)*
- Data abstraction *(A graphical object is defined using the data)*
- Data rendering *(Viewing on a workstation)*
- Data interpretation *(analysis)*
- Feedback *(Perhaps go back to previous phases)*

A problem with the existing CFD software is that it takes a non integrated approach to dealing with the different steps of the CFD process. The grid generation and flow solver programs are involved in the data generation phase. The visualization software is part of the abstraction, rendering and analysis phases. The various programs present the user with different interfaces, and there is little attention paid to the data manipulation and feedback steps. In the current system, large data sets flow from one step to another from disk to ram and back to disk (perhaps from one computer to another), taking on different file formats along the way.

The design criteria for FAST were:

- Minimize the data path in the CFD process
- Provide a consistent user interface
- Allow for quick user feedback
- Provide an extensible software architecture
- Provide a quick path through the CFD process
- Provide libraries and tools so that application modules could be added easily
- To isolate 3D viewing tasks from the application programmer

In order to achieve these design goals FAST has evolved into collection of programs that communicate via Unix sockets with a central hub process that manages a pool a shared memory. A fundamental data type is loaded or generated and stored into shared memory (data generation and manipulation), a collection of programs (modules) operate on data and produce additional data (objects) that are also placed into shared memory (data abstraction). The objects are rendered using the fast viewing system (data rendering). Data is analyzed by additional modules or visual inspection (data analysis). Depending on the results of the analysis the user changes input to any of the previous modules (feedback). In addition there is a collection of libraries and utilities that are used to build the application modules.

The use of shared memory reduces the flow of data in the system. The use of a viewing process relieves the burden of three dimensional interactive viewing from the application programmer. The fact that the fundamental data type(s) reside in shared memory makes it easy to make changes based on the feedback obtained from the analysis phase. Finally the use of FAST libraries and utilities makes it easy to add new modules.

We are aware of other scientific visualization packages and visualization capabilities in existence and/or under development. These include visual programming examples like CONMAN (Silicon Graphics[3]) and AVS (Application Visualization System, Stardent Computer[4]), and other scientific visualization environments like MPGS (Multi-Purpose Graphics System, Cray Research), and the Personal Visualizer (Wavefront), as well as 'scripting' languages like PVWAVE (Precision Visuals), IVIEW (Intelligent Light), and VISAGE (Visual Edge) to name a few. While FAST is built specifically around the research tasks involved in CFD analysis, these other environments and packages typically take a much more generalized approach towards visualization, for the obvious reason that CFD research is a relatively small part of their intended audience. These systems and environments often require a certain level (a 'power' user, visual programmer, or animation /rendering expert) of skill with computer graphics above and beyond the level of the typical CFD scientist. In researching these other more general approaches, we have discovered that the results (data) get handed off at some point to the 'power' user (or perhaps even computer graphics group or expert) and this person (or group) creates the animations, films or videos. FAST is built around a model where the scientist is the first and last person in the data chain and FAST is a toolset for his environment. This is not meant as a criticism of these other approaches, as the need for generalization dictates the need for this other level of user. It is our belief, though, that the techniques used in FAST presented in this paper would also apply and be very useful in the more general environments.

Graphics, CPU, and memory handling performance were key considerations in the FAST design and development process. For graphics, a base-line level of what is commonly termed (but undefined) as "real-time" had to be established and agreed upon as acceptable. This was determined to be a minimum of 3 frames/sec for a typical 10-20 Mbyte problem (techniques used for rendering would determine the problem size in this range). This base line frame rate was determined to be essential in visualization of fluid mechanics for understanding the dynamics of the simulations. For the development platform, the Silicon Graphics 4D220/GTX (16 Mbytes memory) this goal was reached, and we are pleased with the current performance level. The Silicon Graphics 4D320/VGX, has even higher levels of cpu and graphics performance[18], although specific test results do not yet exist.

We have implemented in FAST new techniques and capabilities non-existent in the previous tools and expanded on others. For example, the colormap editing capabilities were enhanced to include banded, spectrum, dynamic, contour, striped, and two-tone function mapping. Surface rendering includes the ability to 'sweep' planes through the data either grid oriented, arbitrarily oriented, or a contour surface (isosurface). Enhanced titling and labelling features include the use of postscript type fonts and symbols, where typeface, font point size, and style can be specified. The animation capability is substantially enhanced beyond what was available in GAS (Graphics Animation System[9]). These enhancements include greater control by allowing the ability to edit scenes, views, and objects. Another capability allows for separate scenes to be rendered in separate windows giving the scientist/user even more flexibility and animation control.

At the time of this printing, the software is in Beta testing at NASA Ames Research Center. The typical workstation environment is a Silicon Graphics 4D/VGX Power Series class machine. The Beta release users currently include approximately a dozen CFD research scientists and application programmers at approximately 250 sites across the country.

FAST ARCHITECTURE

Each separate process communicates through the FAST Hub while managing shared memory and communicating using standard Berkeley UNIX Interprocess Communication (IPC[11]).

Hub

The central process of the FAST environment is the Hub module (Hub, figure 2). The Hub module invokes and shuts down the FAST modules yet its main function is to process requests sent by the modules. These requests might be to allocate a segment of shared memory and return the shared memory id, or to delete a shared memory segment. Since the Hub process is always running as long as FAST is active, the data allocated through the Hub remains accessible even when the original process which requested it is terminated. The Hub module is essentially transparent to the user, in that it has no panels.

Viewer

This is the central module for processing, from the users perspective (Viewer, figure 3). This is where the graphical data pool generated by other modules is managed and interactively viewed. FAST Central, unlike other FAST modules, runs continuously while FAST is up and running. Other modules can be spawned or shut down as they are needed from the Viewer module. In addition Viewer allows object attributes to be set (e.g. transparency, mirroring, line width), scene attributes to be set (e.g. lighting, color map editing, background color), viewing preferences to be set(e.g. toggle axis, mouse axis modes) as well containing the animation control panels. Animator is used to create and record smooth (spline interpolated) keyframe animation sequences.

File I/O

The file i/o module (file i/o, figure 4) loads pre-computed *PLOT3D* type grid, solution, and function files as well as ARCGraph[20] files into FAST's shared memory. It consists of three control panels. The file input panel is used to list file names and information and to load data into shared memory. The data sub-panel displays pertinent information about the previously loaded grids and solutions. The ARCGraph panel is used for handling this type of file input.

CFD Calculator

The CFD Calculator (figure 5) module allows the scientist to attach to the grid and solution data that has been loaded and to calculate a variety of scalar and vector functions for analyzing the computed solution. The Calculator has the appearance and functionality of a real programmable calculator but instead of operating on numbers it operates on *fields* of numbers (scalars) and *fields* of vectors.

Its basic operations (e.g., +, -, MAG, CURL), are applied to entire fields - either component-wise or vector-wise. For example, + applied to two scalar fields will produce a new scalar field of values that are the sums of the corresponding values of the two operand scalar fields. And LOG applied to a vector field will generate a new vector field by taking the logarithm of each component of the corresponding operand vector. In addition to component, scalar and vector binary operators there are also special operations such as GRADIENT, DIVERGENCE, DOT, and CROSS that apply to entire fields and produce new scalar or vector fields.

The scientist can select a range of active solution zones on which to operate and use the CFD Calculator to compute about 100 different built-in CFD scalar and vector functions such as Pressure, Enthalpy, Normalized Helicity, Velocity, and Vorticity [16]. These fields are stored in one of the Calculator's scalar or vector *registers*. The Calculator can then be *programmed* with formulas that operate on these fields and produce new ones using the basic operations already mentioned. The CFD Data Panel is used to copy, move, delete, and display information fields (such as min-max) stored in the Calculator's registers. These features, and others, help make the CFD Calculator an interactive, powerful tool that the CFD scientist can use to compute important quantities for analyzing computed solutions.

SURFER

The SURFace Extractor and Renderer module (figure 6) attaches to grids (loaded by the file i/o Module) and scalar and vector fields (generated by the CFD Calculator) and renders grid surfaces as points, lines, vectors, or polygons. These *grid surface objects* are also stored in shared memory so they can be rendered in the FAST environment. The grid surfaces can show the grid geometry, for example, a lighted, Gouraud [2] shaded polygon surface of the Space Shuttle, or they can display the scalar data as function colored lines or polygons, or vector data as line vectors, vector heads, or polygon vector deformation surfaces (vector heads connected in a surface). Grid surface objects can represent grid geometries, scalar fields, and vector fields.

In addition to changing data types, surface rendering and other attributes, SURFER can sweep through all surfaces in a given grid direction.This creates a dynamic image showing even more features of the flow field.

Titler

The Titler module (figure 7) is used to create high quality Postscript text suitable as titles for images in videos, slides, and movies. Title attributes include font, point size, position, color, drop shadows, and a snap-to-grid feature to make alignment easier. Like other graphical objects, *title objects* are stored in shared memory so they can be added to other scenes. Postscript fonts from other sources may be imported and created titles may be saved for later use.

Isolev

Isolev (figure 8) performs three functions using a single algorithm. One, it draws surfaces of constant value in 3D scalar fields, i.e. isosurfaces. Two, it draws cutting planes function mapped by the scalar field of interest. Cutting planes may be at any angle, and are consistently oriented throughout a multi-zoned grid. Three, it draws vector field deformation surfaces originating at cutting planes or isosurfaces. Iso and deformation surfaces are lighted and smooth shaded. Both isosurfaces and cutting planes may be rendered as dots for improved performance. Interactive grid coarsening is available to improve interactivity. The user may also set up sweeps, where isolev automatically sweeps the isovalue (or cutting plane location) through all possible values, or within a user specified range. This can be used to get a feel for the entire volume. The marching cubes algorithm [Kerlick,13] is used to generate polygons. Level scalar fields are created to generate cutting planes function mapped by the scalar field of interest. Edge crossings, a faster algorithm, is used to generate points. A user selected vector field may be used to draw vectors originating at the crossing points.

Tracer

The tracer module (figure 9) is used to compute particle traces and render them as vectors through the flow field. Tracer attaches to a grid and solution and allows the user to interactively select the point of release or rake[7] from which the traces are computed. The traces can either be computed forward or backward in time as well as allowing the user to selectively save traces. Once traces are saved, a delta time factor may be interactively adjusted through the panel to allow particle trace "cycling".

Topology

The topology module identifies and classifies critical points in a flow field. Critical points are marked with icons which visually identify the class of the point. Traces can be computed at or about these critical points. Topology can find and display vortex cores by examining eigan vectors.

Interactive Visualization Control

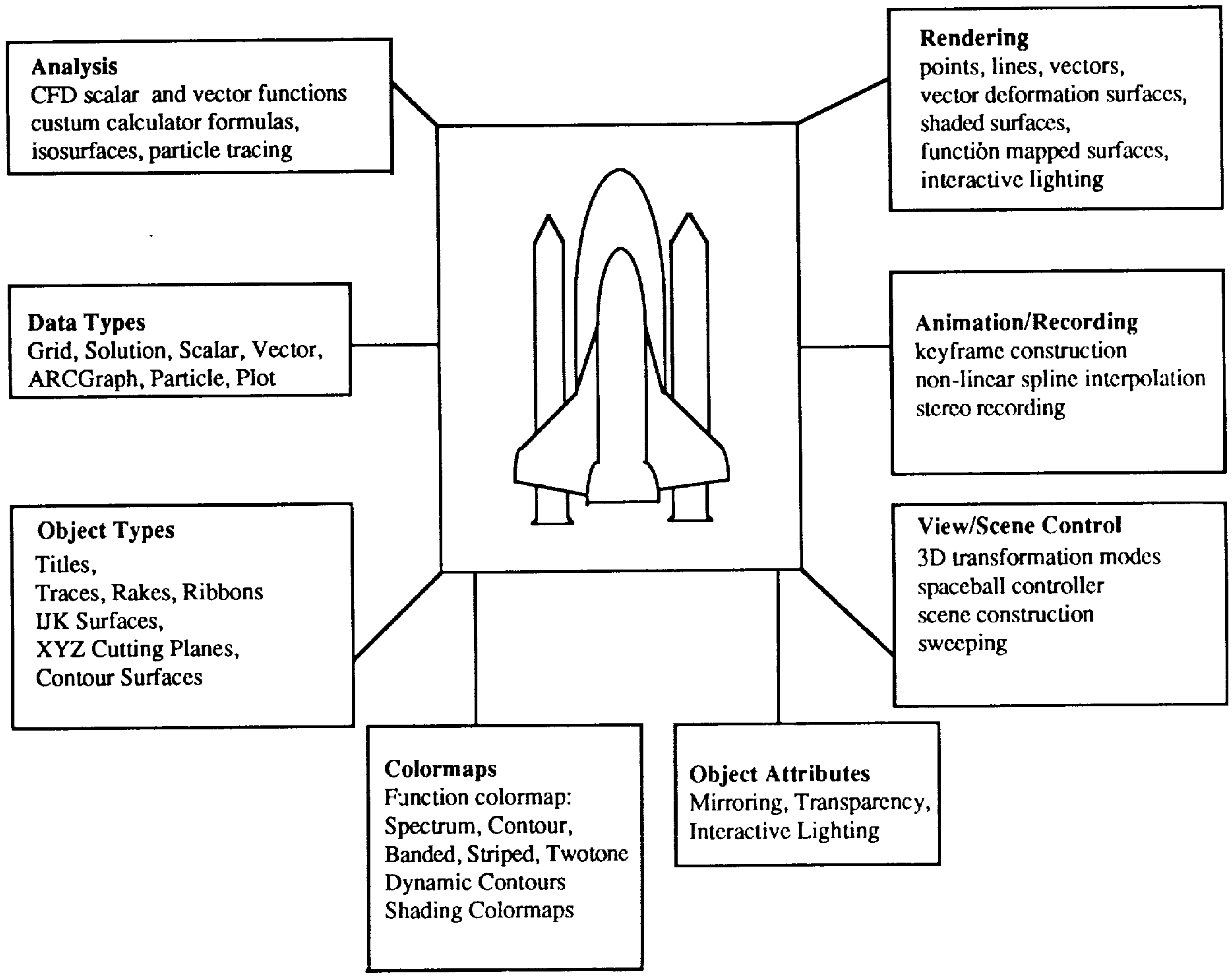


Figure 1, "FAST Interactive Visualization Control"

Interactive Visualization Control

Multi-processed: In figure 3, several modules worked together to generate the scene: Surfer generated the grid surface objects, CFD Calculator computed the scalar and vector fields, Titler was used to generate the text, and Viewer was used for image handling and color map editing. When modules are not needed they can be iconified so they occupy less screen space and CPU resources. Because of this, the FAST environment can be running while other applications are also being used. Alternatively, FAST modules can be terminated without exiting the FAST environment - and this has no effect on their data since it is already in shared memory. Unlike standard dynamic memory, shared memory remains available even after the allocating process is killed. All shared memory segments are removed when FAST is exited via the Quit selection of the Viewer module.

Powerful. The FAST environment contains sophisticated tools such as the CFD Calculator that enable the scientist to analyze computed solutions by examining many relevant "CFD quantities", such as normalized helicity, shock, perturbation velocity, and vorticity . And if these "built-in" functions are not adequate the scientist can program the Calculator to compute customized functions using the rich set of component, scalar, and vector operators. In figure

3, the CFD Calculator was used to compute entropy and pressure scalar fields and a velocity vector field (see "FAST Architecture, CFD Calculator, Page 7).

Flexible. Storing data and graphical objects in shared memory has enabled the complex scenes in figure 3 to be constructed by mixing and matching shared data from any module that is currently plugged into the FAST environment. The figure shows how grid, scalar, and vector data has been combined to generate grid surfaces rendered as grid lines, scalar colored smooth polygon surfaces, and vectors.

Interactive. Surfer provides the ability to interactively alter scene attributes such as coloring the data by a different scalar field, displaying a different vector field, adjusting the legend, normalization, and clipping ranges, or changing rendering and data types. For example, the vector field can be rendered as a Gouraud shaded, lighted, vector deformation surface. With the looping option turned on Surfer will sweep through all data in the current grid direction - providing a dynamic visualization ability. And while this is happening the scientist (from Viewer) can transform (e.g., rotate, or zoom) all or part of the scene or use the color map editor to adjust the function color mapping by inserting, deleting, and changing colors, or selecting a different colormap types such as *Spectrum*, *Contour*, *Striped*, *Twotone*, or *Banded*.

IPC and Shared Memory Implementation

It was decided that an interprocess communication (IPC) package must be implemented to allow FAST to operate as a modular environment where resources could be shared among different machines as well as a single host. Specifically, Unix System V shared memory facilities are used to allow each process (module) to access the environment's data, while the Berkeley IPC package's implementation of Internet domain stream sockets allows for the coordination of this data.

As each module is executed by the FAST hub, it must immediately establish a two-way communication channel between it and the hub. Because an Internet domain address consists of a machine network address and a port number, these two values are used in establishing this connection. The following command is therefore executed at the beginning of a module's main routine:

```
socket_establish_and_accept (hub_host, hub_port, &rsock, &wsock);
```

This does the following:

1) create a socket from which to read
2) determine a local port and listen on it
3) create a writeable socket and establish a connection to the hub (using the hub's hostname and port number which came in as arguments)
4) now send the port number to the hub and
5) accept a connection from the hub

At the same time, the hub process executes this statement:

```
socket_accept_and_establish( sock, module_host, &wsock );
```

which does the following:

1) create one socket from which to read from all modules
2) accept a connection from the next module
3) read in the module's port number
4) create a writable socket and connect this socket to the module

After a two-way connection has been established, both the hub and the module are left with two socket descriptors each. These are used exactly as a file descriptor is used, one for writing (wsock) and the other for reading (rsock). The hub actually stores these descriptors along with other pertinent information, such as module status, in an array of structures - one structure for each module.

The modules specified for inclusion in the FAST environment are specifically listed in a "run command" file called $HOME/.fastrc. Also included within this file is information about initial placement of a module's main panels, the name of the host where the module resides, and the complete path name of the particular module.

Once a module has been executed by the FAST hub using the Unix system(3) call and the communication channels have been established, the hub enters a loop where it waits on a request from any of the active modules to perform some sort of action. The hub process uses the Unix select(2) call to examine all available read socket file descriptors to determine if they are ready for reading. This appears as follows:

```
while (continue_looping) {
        for each module
                load read socket id into fds, file descriptor structure
        select (fds, 1,000,000 seconds)  i.e. pause here until a request is detected
        communication is detected ... determine from which module
        read up the request from that module
        process request
} end while
```

Information sent between a module and the hub (and vice-versa) is always preceded by a standard sized structure which contains, the command and four information fields. The necessary information, if any, is then written back to the module, and the flow control takes the hub back up to the point where it can again wait for a request.

One example of a request that a module might make would be the allocation of memory which may eventually be used by another module. It must first send a request to the hub to do this. The hub then allocates the memory as a shared memory segment and retrieves the shared memory identifier associated with this segment. This identifier is then stored by the hub in a data structure possibly to be accessed by another module at a later time. Finally this identifier is sent back to the module so that it may attach the shared data to it's virtual memory address space.

At any time that a different module would like to access this data, a request is similarly sent to the hub to retrieve the shared memory identifiers so that it too may attach to the data.

A consequence of using shared memory instead of standard dynamic memory is that dynamic data structures such as linked list nodes no longer have a *pointer* to the next node but rather the *shared memory id* of the next (and current) node. And this shared memory id must be explicitly *attached* to and *detached* from whenever the structure is traversed.

The FAST environment contains several lists of this form: a list of grids, a list of solutions, a set of scalar and vector lists (one for each register of the CFD Calculator), and a list of graphical objects. A typical request that a module would make of the Hub is to gain access to a particular list node, for example, a node from one of the CFD Calculator's vector register lists. This would involve setting up the fast_infobuf with the appropriate information about the request, writing it to the Hub, reading the node's shared memory id from the Hub, and attaching to generate a virtual address for the requesting module process. The Hub process detects the socket write in its main event loop and executes a socket read and calls the function process_request() to handle the module's request:

Modules that generate data to be shared must: 1) change low-level usage of pointers to shared memory ids, 2) alter management routines to explicitly attach and detach in addition to allocate and deallocate, 3) provide a library of routines that modules can link with that provide access to the actual data stored in these structures, 4) provide a library of routines that the Hub can use to create and destroy these structures (recall that the Hub is the single process that does all shared memory allocation and deallocation).

Graphical objects are also shared which means the structures that define them must reside in shared memory. Note that part of this structure references the shared memory ids of the grid, scalar and other data needed to draw a grid surface object. The routine draw_grid_surface() accepts this structure and draws it. This routine is part of the viewing library which is linked to every FAST graphical module so they can all include grid surface objects in their scenes.

Using shared memory and sockets, FAST is able to quickly and easily share all the data used within the environment. Even though shared memory can not yet be shared over different machines as it is on a single host, FAST has been designed with that feature in mind. When indeed we can accomplish this, the ultimate power of FAST can be realized.

DISCUSSION

For an existing SGI visualization application to be converted into a FAST module:

- Command line arguments must be used to establish window location and Hub communication - and nothing else.
- Periodically, each module must check for exit command IPC from the Hub. This is usually done once each time through the main event loop.
- Standard input should not be used.
- Standard output should be used sparingly for status and error messages.
- The colormap must be used according to FAST conventions. FAST library functions must be used to get color indexes for drawing. A few indexes are reserved for modules to create their own colors, but most of the colormap is only modified via the FAST COLORMAP module.
- Grid, vector and scalar field data must be accessed via FAST shared memory.
- The panel library should be used for menus, buttons, sliders, etc.
- The panel library's nap time or blocking should be turned on when waiting for user input to avoid excessive context switching.
- The application's drawing code must be integrated into the viewing library so that it's visualizations can appear in all modules.
- The data needed to draw must be placed in shared memory and made available to the viewing mode

There are several advantages to integrating applications into FAST as modules. These advantages include:

- Shared memory speeds which allow users to interactively view their data from several modules without long disk IO delays.
- Access to CFD Calculator generated vector and scalar fields.
- Precalculated min and max for grids, vector and scalar fields. This reduces the time needed to access data in many cases.
- Sophisticated colormap manipulation using the FAST COLORMAP module.
- Integration of visualizations created by several modules into a single scene.
- Trivial integration of visualizations into animations.
- Interactive access to most of the generic capabilities of the SGI graphics hardware, e.g. rot-tran-scale, using the viewing library panels.
- Other synergistic effects of multiple modules accessing the same data.
- New applications can be built quickly since many functions are be made available by existing FAST modules and libraries.

There are also some disadvantages, of course. These include:

- Time to learn to use the FAST libraries and intermodule communications as well as to keep up with future changes.
- Performance overhead due to multiple processes busy waiting.

Future plans for FAST include the capability for use across high speed LANs for 'smart' distribution of processing. Compute intensive modules could be distributed or broken up into components that communicate over these networks, or perhaps memory could be shared across systems.

As flow solvers become fully integrated, and interactive 3-d grid generation becomes a reality, FAST will continue to offer more effective visualizations of computational aerodynamics in all aspects of fluid flow simulations.

N92-22439

A NEAR-REAL-TIME FULL-PARALLAX HOLOGRAPHIC DISPLAY FOR REMOTE OPERATIONS

Helene P. Iavecchia*
Computer Sciences Corporation
Moorestown, NJ

Lloyd Huff
University of Dayton Research Institute
Dayton, OH

Neville I. Marzwell
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA

ABSTRACT

A near-real-time, full-parallax holographic display system was developed that has the potential to provide a 3-D display for remote handling operations in hazardous environments. The major components of the system are: (1) a stack of three spatial light modulators (3-SLM stack) which serves as the object source of the hologram; (2) a near-real-time holographic recording material (such as thermoplastic or photopolymer); and (3) an optical system for relaying SLM images to the holographic recording material and to the observer for viewing.

I. INTRODUCTION

Over the past 30 years, holography has made substantial contributions to several fields. Holography is a core technology in nondestructive testing to find hidden defects beneath surfaces, in interferometry to measure small mechanical displacements, and in optical pattern recognition. Three-dimensional (3-D) holographic displays are common in art and advertising and embossed holograms provide enhanced security for credit cards, entertainment tickets, record albums, and clothing labels. Holographic displays, however, are typically time-consuming to initially generate and are usually static. This paper describes the prototype development of the Holographic Enhanced Remote Sensing System (HERSS) for generation of near-real-time 3-D snapshots of remote equipment and objects[1]. This effort was conducted over the period 1988 through 1990.

II. HERSS GOALS

The key goal of the HERSS program was to generate a near-real-time, full-parallax (i.e., having both horizontal and vertical parallax) holographic snapshot of remote objects using data collected by a conventional imaging system. Specifically, the image data would contain the surface points of the remote objects and could be collected, for example, using a laser range scanner. The goal of the prototype system was to generate a 3-D snapshot of the remote image data in 30 seconds or less. While 30 seconds may seem a relatively long interval, this image update period was still deemed adequate for the intended application and is considered to be near real time in the context of remote operations. Furthermore, generating a *full-parallax* holographic image in this time frame is an ambitious goal which challenges the performance capabilities of currently available components in critical parts of the system.

---

* H. P. Iavecchia was with Analytics, Inc., Willow Grove, PA when this work was conducted.

REFERENCES

1. P. Buning et al., "Flow Visualization of CFD Using Graphics Workstations", AIAA 87-1180, *Proc. 8th Computational Fluid Dynamics Conf., June 9-11,1987.*
2. J. Foley and A. Van Dam, *Fundamentals of Interactive Computer Graphics*, Addison Wesley, 1982
3. P. Haeberli, *"ConMan: A Visual Programming Language for Interactive Graphics,"* SIGGRAPH Proceedings, Volume 22, Number 4, SIGGRAPH August 1988
4. C. Upson, et al., *"The Application Visualization System (AVS): A Computational Environment for Scientific Visualization"* IEEE Computer Graphics and Applications, July 1989
5. J. Helman, L. Hesselink, *" Representation andDisplay of Vector Field Topology in Fluid Flow Data Sets"*, IEEE Computer, August 1989
6. G. Bancroft, *"Scientific Visualization in Computational Aerodynamics at NASA Ames Research Center*, IEEE Computer, August 1989
7. P. Walatka, P. Buning, *PLOT3D Users Manual Version 3.6* NASA TM101067,1989
8. T. Plessel, *SURF Users Manual*, NASA Ames Research Center, Code RFW, 1988
9. T. Plessel, *GAS Users Manual*, NASA Ames Research Center,Code RFW, 1988
10. S. Leffler, et al.,*"The Design and Implementationof the 4.3 BSD UNIX Operating System"*, Addison-Wesley,1989
11. K. Haviland, B. Salama,*"UNIX System Programming"*, Addison-Wesley,1987
12. *Stellix Programmers Guide* , Chapters 15-17,Stellar Computer Inc.,1988
13. Lorenson, W.E., and Cline, H.E., *"Marching Cubes: a High Resolution 3D Surface Construction Algorithm,"* Computer Graphics, Vol 2.1, No. 4 July 1987, pp;. 163-169.
14. D. Tristram, P.Walatka, E. Raible *"Panel Library Programmers Manual, Version 9.5"*, NASA Ames Report RNR-90-006
15. F. White, *"Fluid Mechanics"*, McGraw-Hill 1979
16. D. Anderson, *"Compuational Fluid Mechanics and Heat Transfer"*, Mcgraw-Hill 1984
17. L.M. Milne, *"Theoretical Aerodynamics"*, Dover Press 1973
18. T. Lasinski, "*Second Generation Graphics Workstations, Request For Proposals*", RFP2-33491, January 1989
19. S. Rogers, *"Distributed Interactive Graphics Applications in Computational Fluid Dynamics"* , Interanational Journal of Supercomputing Applications, Vol 1, No. 4, Winter 1987
20. E. Hibbard, *"ARCGRAPH (Ames Research Center Graphics Metafile) Manual"*, NASA Ames Research Center

Certain subgoals were also established for HERSS development including:

- Map the surface points occupying a 152 x 152 x 152 mm (6 x 6 x 6 in) cube at the remote work site onto a holographic volume of the same size;
- Provide a depth resolution in the 3-D display of 2 to 3 mm (0.08 to 0.1 in), and;
- Provide a capability to overlay graphics onto the hologram.

III. APPROACH

The HERSS concept for generating a holographic snapshot involves four basic steps. First, the numerical representation of object surface points is collected. Second, the numerical data base is "sliced" into 2-D depth planes with a finite thickness. If the slice thickness is 3 mm, then each 2-D depth plane contains the surface points of the objects in a unique 152 x 152 x 3 mm (6 x 6 x 0.1 in) region. That is, all surface points in a 3 mm region of depth are compressed into one 2-D plane. Third, each 2-D plane is sequentially transmitted to a computer-addressable spatial light modulator (SLM). The SLM acts as the "coherent" object source for the hologram. Each SLM image is exposed on a near-real-time holographic recording material (HRM), such as thermoplastic[2] or photopolymer[3], using a plane-by-plane multiple exposure process. Images are sequentially transmitted to the SLM for exposure until all depth planes are recorded. Finally, the HRM is developed through a heating process for the thermoplastic medium or through a UV bath for the photopolymer.

When the HERSS project was initiated, the viability of using an SLM as a holographic object source had been demonstrated with the production of holographic stereograms on photographic film[4]. Furthermore, the viability of multiple-exposure holography had also been demonstrated with the generation of a 3-D view of CAT scan data by multiply exposing superimposed CAT cross-sections onto standard silver halide film[5]. No published reports, however, were available that documented using an SLM as a holographic object source for a near-real-time HRM. Furthermore, no reports documented multiple exposure of thermoplastic or photopolymer. It was unknown how many exposures each of these HRMs could store before image quality would be seriously degraded.

IV. BASELINE CONFIGURATION

A. Optical Configuration

The baseline optical system, as shown in Figure 1, was designed to generate an image-plane transmission hologram. For baseline testing, an nView SLM was selected as the holographic object source and a thermoplastic camera developed by Newport Corporation (based on an original design by Honeywell) was selected as the near-real-time HRM. The nView SLM had an addressable 152 x 152 mm (6 x 6 in) area. Because the Newport thermoplastic material was only available in a 38 mm format, this created the need for magnifying optics to enlarge the holographic image for the observer. Thus, the optical configuration served three functions: (1) direct an expanded object beam to the SLM and an expanded reference beam to the HRM; (2) relay the SLM object beam image to the HRM; and (3) magnify the HRM image for viewing.

B. Plane-by-Plane Exposure

The plane-by-plane multiple exposure technique initially employed by HERSS is illustrated in Figure 2. Laser light transmitted through open SLM-cells outlines the shape of the remote object(s) for a particular depth slice. A projection lens is used to relay the SLM image to the HRM for exposure. The SLM is in a fixed position while the projection lens is mounted on a sliding table that has a 152 mm (6 in) travel. When the first depth plane is exposed, the table is positioned at one end of its travel. After each exposure, the table is moved to the next depth plane location. Unique depth plane images are sequentially transmitted to the SLM for exposure on the thermoplastic. This move-display-expose cycle is repeated until the entire object volume is recorded. When the exposure process is complete, the material is developed and a single holographic frame is viewable.

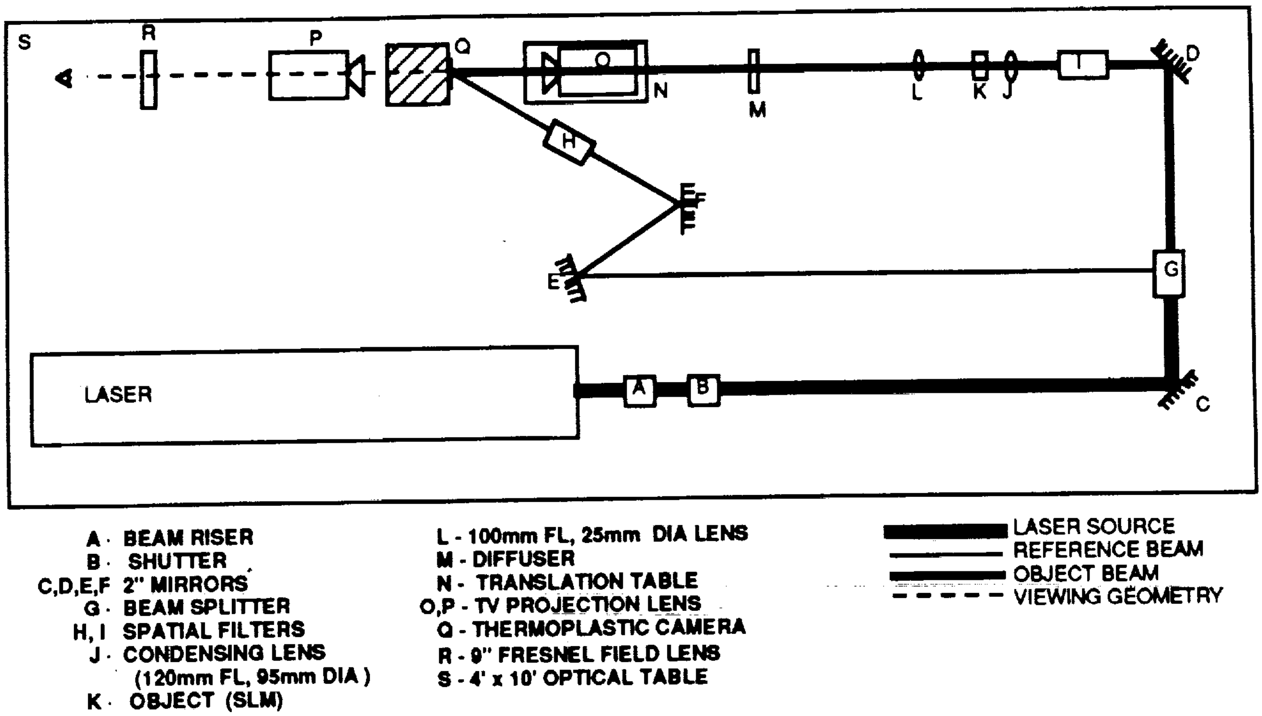


Fig. 1. Baseline Optical Configuration. The schematic portrays the optical paths of the object and reference beams as well as the viewing geometry of the system. The object beam passes through the SLM (K) on its path to the recording material (Q). The reference beam is directed to (Q) where it interferes with the object beam. This interference pattern stores the intensity and phase information necessary to reconstruct the 3-D image for the viewer.

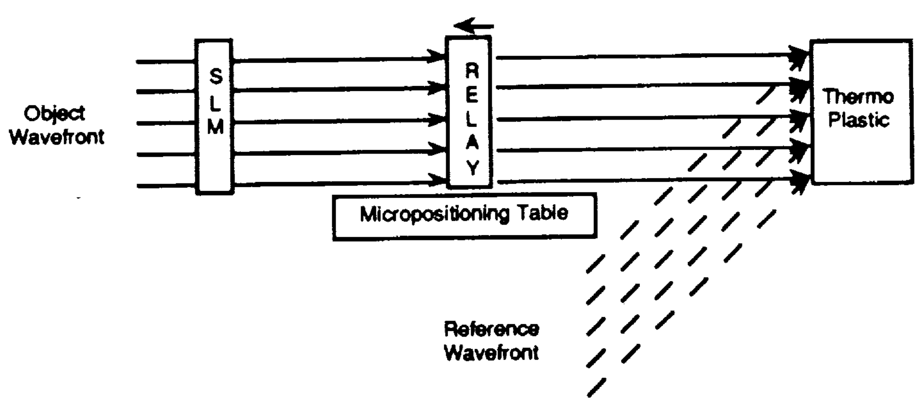


Fig. 2. Baseline Plane-by-Plane Exposure Method

C. Results

Early experimentation with the baseline configuration yielded two major successes regarding holographic image generation and quality including:

- Generation of a hologram was possible while the relay lens was mounted on a motor-driven micropositioning table. Initially there was some concern that vibration transmitted from the stepper motor to the table would negatively affect the SLM light pattern, resulting in a failed hologram.
- Twenty depth planes (using a patterned glass target, not an SLM image as the object source) could be superimposed through multiple exposure onto the thermoplastic without serious degradation of image quality. (The recordings were made using a K3 ratio of 1 and an exposure energy of 7 $\mu J/cm^2$ with $1/n^{th}$ the exposure energy for any single depth plane.)

However, drawbacks regarding image quality were also noted including:

- A sharp holographic image could not be generated using the nView SLM as the holographic object source. This was attributed to a low contrast level of the SLM. Considerable light leakage through closed cell areas was evident.
- The holographic image was distorted by the optical components that relayed the SLM image onto the HRM as well as by optical components that magnified the image for viewing.

Regarding HERSS goals, the following conclusions were drawn:

- To meet the display-depth-resolution goal of 2 to 3 mm (0.08 to 1.0 in), at least 51 depth planes would need to be recorded in the 152 mm (6 in) holographic depth. With 20 superimposed depth-planes, a depth resolution of only 7.5 mm (0.3 in) could be achieved.
- Even though a 152 mm (6 in) holographic image depth could easily be achieved, the optically-relayed image was laterally limited to 114 mm (4.5 in). This was the direct result of a relatively small Newport thermoplastic film format. Optics to minify and relay the SLM-image (152 x 152 mm) to the thermoplastic (38 x 38 mm) as well as optics to magnify and relay the thermoplastic image to the observer could only practically meet a threefold minification or magnification. A fourfold factor would require very low-f/# optics that are both difficult to produce and are prohibitively expensive. A larger aperture film-based thermoplastic system was evaluated; unfortunately, the efficiency and reliability of the system were inadequate.
- To record 51 image planes using the plane-by-plane exposure technique would require about 64 seconds for exposure and 30 seconds for development. Each exposure consists of moving the micropositioning table to the next depth position (250 msec) and allowing time for table vibrations induced by the motor to dissipate (1000 msec). While the 30-second development time could be reduced to less than 100 msec via modification of the electronic components heating the thermoplastic[2], the plane-by-plane exposure method itself is still a time-consuming process.

V. REVISED CONFIGURATION

A phase conjugate optical configuration was devised to circumvent the image distortion produced by the baseline optical system. A multiplane exposure technique was also devised to (1) increase the number of image planes recorded in a holographic snapshot and (2) reduce exposure time. For the revised configuration testing, the nView SLM was replaced by three Sharp SLMs which were extracted from a SharpVision Projection System. The Sharp SLMs are monochromatic TFT-based units that are 2.4 inches wide by 1.8 inches high, having a resolution of 384 x 240 pixels and a 30:1 contrast ratio. Testing was continued using Newport thermoplastic as the HRM. Some testing was also conducted using Du Pont HRF-700 photopolymer as the HRM.

A. Phase Conjugate Optical Configuration

The revised optical configuration, as shown in Figure 3, was based on the principles of phase conjugation and reverse ray tracing[6,7]. In a phase conjugate optical system, the optics used to generate the hologram are the same optics used to view the hologram. Specifically, during hologram generation, the object and reference beam wavefronts interfere at the front face of the holographic recording material, as is required for generation of a transmission hologram. During viewing, the reference wavefront is transmitted through the rear of the recording material and travels in a reverse path through the optical system to reconstruct the original object wavefront. Image distortions are still induced by the optics during hologram generation. However, these distortions are removed during the phase conjugate reconstruction when the light diffracted by the hologram is transmitted through the optical system in a reverse path to produce the 3-D image that the observer views.

The phase conjugate configuration also provides a one-to-one correspondence between the size of the SLM image and the size of the recorded holographic image. This means that the lateral dimensions of the holographic image are equal to the lateral dimensions of the SLM. Thus, magnifying optics to enlarge the image for viewing can be eliminated.

B. Multiplane Exposure

To circumvent the multiple exposure limitation of the thermoplastic, a multiplane exposure method was devised as illustrated in Figure 4. Basically, the method uses a stack of three SLMs. Each SLM in the stack is slightly separated in the axial plane to represent the real distance between image slices. During the multiplane exposure process, three depth plane images are simultaneously exposed on the thermoplastic instead of only one depth plane at a time. The revised technique has the potential to increase the number of image planes recorded in a holographic frame because, with only 17 exposures, 51 image planes could be recorded.

In addition to increasing the information content within a holographic frame, another key benefit of this technique is the potential to reduce the time to generate a single holographic frame. If 51 images are recorded in 17 exposures instead of 51 exposures, then 34 micropositioning table moves can be eliminated as well as the concomitant settle time for each table move. This provides a significant savings of time considering table move and settle times of 250 msec and 1000 msec, respectively.

C. Results

Key successes of the phase conjugate configuration and the multiplane exposure technique included:

- Distortion-free holographic imaging with the phase conjugate reconstruction.
- Image quality comparable to, if not better than, the baseline.
- A 95% increase in the number of image planes that could be recorded on the thermoplastic HRM. Specifically, 39 depth planes could be superimposed with 13 exposures of the 3-SLM stack before image quality was seriously degraded. With the baseline plane-by-plane exposure technique, only 20 depth planes could be superimposed before quality degraded. This successful recording of 39 depth planes demonstrates the viability of the stacking method. (The recordings were made using a K ratio of 1 and an exposure energy of 7 $\mu J/cm^2$ with $1/n^{th}$ the exposure energy for n exposures of the stack.)
- The resolution and contrast of the Sharp SLMs provided acceptable imaging.
- Experimentation with alternative SLM-addressing schemes revealed a flexibility in selecting how surface shapes would be represented. Both a "bright on dark" and a "dark on bright" holographic image were generated. A"bright on dark" image was created if *open* SLM pixels represented surface points. If *closed* SLM pixels represented surface points, the light transmitted through the SLM to the HRM resulted in a "dark on bright" image. The latter option is relatively simple to implement — pixels containing shape information are closed while all other pixels are open. The former option requires that pixels containing shape information are open while others are closed. However, SLM pixels in the stack that are in front of or behind open image-containing pixels must also be opened to allow light to be transmitted. Even though more difficult to implement, the "bright on dark" technique was adopted for prototype testing because the non-image light in the "dark on bright" technique could possibly degrade the multiple exposure recording capability of the film.

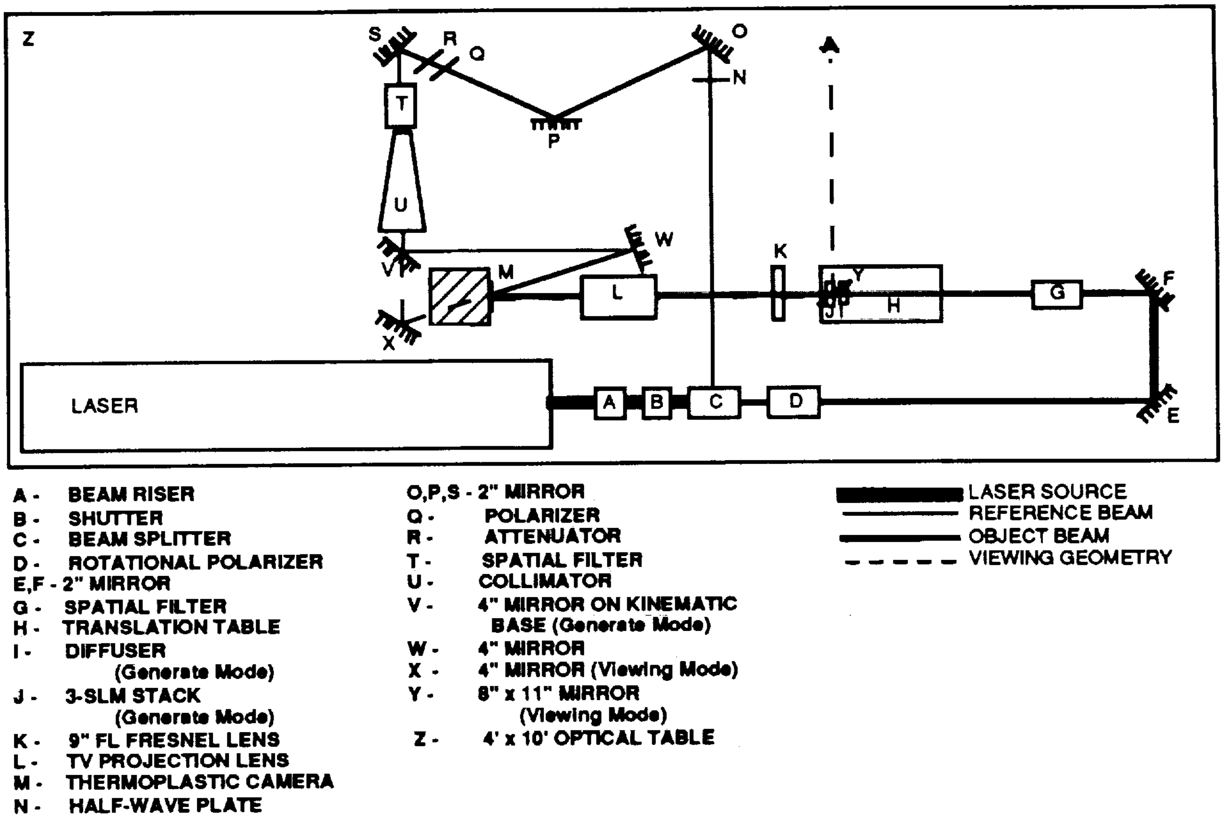


Fig. 3. Phase Conjugate Optical Configuration. The schematic portrays the optical paths of the object and reference beams and the viewing geometry of the system. The object beam passes through the 3-SLM stack (J) on its path to the recording material (M). The reference beam is also directed to (M) where it interferes with the object beam during hologram recording. During reconstruction, the object beam is blocked by removing the mirror on the kinematic mount (V). This allows the reference beam to reflect from mirror (X) onto the rear of the holographic recording plate. The diffracted light then passes back through the optics in the direction of the observer. During viewing mode, the 3-SLM stack (J) and diffuser (I) are removed and a mirror (Y), oriented at a 45° angle from the optical path, is mounted on the micropositioning table (H) in place of the SLM stack.

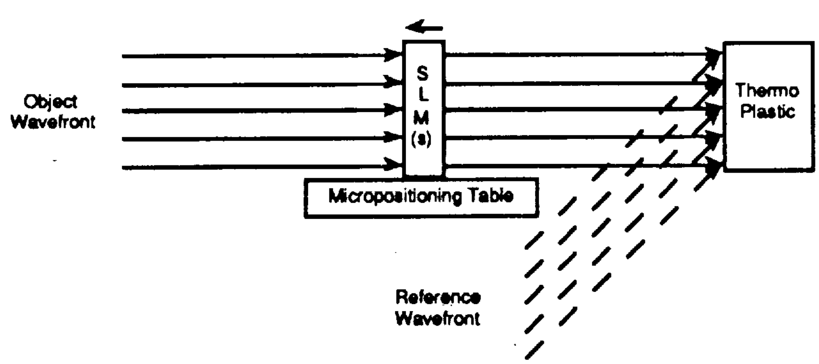


Fig. 4. Multiplane Exposure Method

Drawbacks regarding image quality were also noted:

- Concerning the thermoplastic material, there was a brightness variation in the holographic image. Brightness varied as a function of viewing angle and was particularly pronounced in the horizontal plane. This image brightness variation with viewing angle was attributed to the diffraction efficiency rolloff of the thermoplastic film with holographic fringe spatial frequency variation above and below the optimum value. Object rays that are recorded at the optimum spatial frequency of 800 line pairs per millimeter (lpm) will exhibit the optimum diffraction efficiency and will be bright and have high contrast. In our configuration, this optimum spatial frequency occurred for normally incident object rays. Rays that are either to the left or right of the normal will be recorded with higher or lower spatial frequency, a corresponding lower diffraction efficiency, and a corresponding shift in brightness.
- A maximum viewing field angle of only 15-20 degrees could be achieved with the thermoplastic camera. This was also attributed to the limited recording resolution of the thermoplastic.
- Concerning multiplane exposure, depth positions were clearly discernable from one 3-SLM stack position to the next. However, it was not possible to discern the difference in depth within a stack. It is likely that this effect is attributable to the relatively small field-of-view of a single pixel using the "bright on dark" SLM-addressing technique (i.e., only one pixel in front of and behind an image-containing pixel is opened).
- Concerning use of the 3-SLM stack for multiplane exposure, a faint Moire pattern could be seen in the background of the holograms when the observer's head moved through the field of view. This problem is attributed to a slight misalignment in the SLM stack. Each SLM has a periodic pattern of thin opaque (5-10 micron) electronic structures that separate the pixels. Superposition of the nonaligned electronic structures causes the Moire pattern resulting in some obscuration of the desired intensity pattern.

Regarding HERSS goals, the following conclusions were drawn:

- With 39 superimposed depth-planes, a depth resolution of only 3.9 mm (0.15 in) could be achieved. This is still short of the desired display-depth-resolution goal of 2 to 3 mm (0.08 to 1.0 in) or at least 51 depth planes in the 152 mm (6 in) holographic depth.
- The lateral dimensions of the holographic image was limited to the lateral dimensions of the Sharp SLM or 61 x 46 mm (2.4 x 1.8 in). While the baseline nView SLM could meet the size criteria, image quality was very poor. The Sharp SLM size limitation is expected to be alleviated when larger format SLMs, currently under development, are released. The development of a Sharp SLM with a 5.5 inch diagonal and a 100:1 contrast ratio is underway[8].
- To expose 51 image planes using the multiplane exposure technique would require about 21 seconds using the thermoplastic material.

D. Testing and Results using Photopolymer as the HRM

Because of the relatively low recording resolution and small format of the thermoplastic, hologram generation using a newly released Du Pont photopolymer (HRF-700 Series) was undertaken. The photopolymer is a volume recording material with a resolution of over 4000 lpm. This means that the photopolymer can store much more information than thermoplastic which has only 800 lpm. An improved viewing field angle is also possible with the higher recording resolution. Furthermore, the photopolymer is available in sheet sizes 8.5 by 11 inches allowing an ideal HRM size of 51 x 51 mm (2 x 2 in). As previously mentioned, affordable optics to minify the SLM image onto the HRM can only practically achieve a threefold minification from 152 x 152 mm (6 x 6 in) to 51 x 51 mm (2 x 2 in). It is also noteworthy that the minification of the SLM image onto a smaller format HRM is critical for minimizing exposure time. Given a fixed laser power, the light intensity per unit area on the HRM increases as the size of the HRM decreases. Thus, as HRM size decreases, exposure time is also decreased.

Due to time constraints, experimentation with the photopolymer was limited to testing with the baseline plane-by-plane exposure technique. Testing focused on determining the maximum number of image planes that could be recorded. The most dramatic testing result was that a high quality image could be produced which stored 40 image planes. This was a 100% increase over the capability of the thermoplastic using the plane-by-plane exposure technique. Furthermore, given additional research, the experimental team was confident that an even greater number of depth planes could be stored on the photopolymer. However, as the number of exposures increase and the energy per exposure decreases (i.e., with each plane assigned 1/nth of the total exposure requirement), it may be necessary to first energize the material to activate the polymerization process.

A drawback of the photopolymer, however, is a slower photospeed compared to the thermoplastic. Exposure time for this material is approximately 24 - 30 seconds using an argon-ion laser (with one watt energy on a single line); development time is instantaneous with the UV bath. To record 51 image planes on photopolymer using the plane-by-plane technique would require about 88 to 94 seconds. Considering a system employing the 3-SLM exposure method, a holographic frame with 51 image planes can theoretically be recorded in approximately 45 seconds (21 seconds for the 17 move-settle cycles plus 24 seconds for exposure).

VI. CONCLUSIONS

A full-parallax holographic display system was developed that can generate 3-D snapshots of a remote work site using data gathered by a conventional imaging system (e.g., a laser range finder). The major components used to generate a holographic snapshot are: (1) a stack of three spatial light modulators (3-SLM stack) which serves as the object source of the hologram; (2) a near-real time holographic recording material (e.g., thermoplastic or photopolymer); and (3) an optical system for relaying SLM images to the holographic recording material and to the observer for viewing.

The viability of the HERSS system concept was demonstrated during prototype development and testing. However, as with any technology development effort, there is a complex set of design tradeoffs that affect program goals. For HERSS, a major tradeoff exists between information content of the display and display update rate. Simply stated, as the number of recorded depth planes increases, recording time increases. Another major tradeoff exists between image quality and display update time. For example, while the recording resolution of photopolymer (at least 4000 lpm) is far superior to that of the thermoplastic material (800 lpm), the energy requirements to expose photopolymer (at least 10 mJ/cm^2) is three orders of magnitude greater than thermoplastic (7 μJ/cm^2) resulting in a slower update rate.

Individual component improvements can lessen the impact of these tradeoffs. For example, to achieve a high quality image with a relatively fast display update rate, a thermoplastic material with greater resolution or a photopolymer with a faster photospeed is desirable. Furthermore, to achieve the goal of a 152 x 152 x 152 (6 x 6 x 6 in) holographic image volume, a high-contrast SLM with lateral dimensions of 152 x 152 mm (6 x 6 in) is needed.

Future experimental efforts that use the current HERSS configuration and components can also be conducted to improve image quality and a faster display update rate. These efforts would include:

- Determine the maximum number of image planes that can be recorded on the photopolymer using the 3-SLM multiplane exposure technique. With the 3-SLM stack, it is possible that as many as 80 image planes can be recorded, thereby exceeding the HERSS depth resolution goals. Use of the 3-SLM stack also significantly reduces hologram generation time.
- Investigate the optimum SLM-pixel-addressing scheme for maximum perception of depth in the display (e.g., creation of the dark-on-bright multiply-exposed hologram).
- Investigate methods to reduce exposure time (e.g., determine the minimum settle-time for each move-display-expose cycle. It is very likely that the current 1-second settle time can be substantially reduced.)

Finally, it is important to conduct psychophysical experimentation to determine what configuration options will result in the most accurate perception of discrete depth planes.

Full-parallax 3-D displays have the potential to provide safer and perhaps more timely remote work operations. While the design challenges associated with developing such displays are demanding, the effort is warranted. It is only with a full-parallax display that an operator can correctly perceive the relative orientation and location of objects at any viewing angle. This is a critical requirement for control tasks in hazardous environments. Finally, the existence of such displays may also provide the opportunity to bring remote equipment operation to close-in tasks which now require direct human control.

The work was performed as part of a Phase II research effort awarded by the National Aeronautics and Space Administration to Analytics, Inc., under the Small Business Innovation Research Program, contract NAS7-1036, monitored by the Jet Propulsion Laboratory. The authors also acknowledge the efforts of other members of the HERSS engineering team in the design and development of the holographic display system including Edwin S. Gaynor, Thomas J. Janiszewski, Kristina M. Johnson, Sarvesh Mathur, William T. Rhodes, and Edward H. Rothenheber.

ACKNOWLEDGMENT

The research described in this paper was partially carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

REFERENCES

1. H.P. Iavecchia, E.S. Gaynor, L. Huff, T. Janiszewski, W.T. Rhodes, and E.H. Rothenheber. Holographic Enhanced Remote Sensing System (HERSS). Analytics Final Report 2168. Willow Grove, PA: Analytics, Inc. (1990).
2. T.C. Lee, N.I. Marzwell, F.M. Schmit, and O.N. Tufte. Appl. Opt. **17**, 2802 (1978).
3. A.M. Weber, W.K. Smothers, T.J. Trout, and D.J. Mickish. Hologram recording in Du Pont's new photopolymer materials. SPIE OE/Lase Proceedings on Practical Holography IV (1212-04), Los Angeles, CA, (Jan 14-19,1990).
4. J.R. Andrews, B. Tuttle, M. Rainsdon, R. Damm, K. Thomas, and W.E. Haas. SPIE Proceedings on Three-dimensional Imaging and Remote Sensing Imaging, **902**, 92 (1988).
5. K.M. Johnson, M. Armstrong, L. Hesselink, and J.W. Goodman, Appl. Opt. **24** (1985).
6. R.J. Collier, C.B. Burckhardt, and L.H. Lin. *Optical Holography* (Academic, New York, 1971).
7. H.M. Smith. *Principles of Holography* (Wiley, New York, 1975).
8. M. Adachi, T. Matsumoto, N. Nagashima, T. Hishida, H. Morimoto, S. Yasuda, M. Ishii, and K. Awane. A high-resolution TFT-LCD for a high-definition projection TV. Proceedings of the Society for Information Display (May 1990).

ELECTRONICS

(Session A5/Room B4)

Tuesday December 3, 1991

- Nonvolatile, High-Density, High-Speed, Magnet-Hall Effect Random Access Memory
- Analog VLSI Neural Network Integrated Circuits
- Monolithic Microwave Integrated Circuit Water Vapor Radiometer
- A Noncontacting Waveguide Backshort for Millimeter and Submillimeter Wave Frequencies

N92-22440

NON-VOLATILE, HIGH DENSITY, HIGH SPEED, MICROMAGNET-HALL EFFECT RANDOM ACCESS MEMORY (MHRAM)

Jiin C. Wu
Center for Space Microelectronics Technology
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California 91109

Romney R. Katti
Center for Space Microelectronics Technology
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California 91109

Henry L. Stadler
Center for Space Microelectronics Technology
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California 91109

ABSTRACT

The micromagnet-Hall effect random access memory (MHRAM) has the potential of replacing ROMs, EPROMs, EEPROMs and SRAMs because of its ability to achieve non-volatility, radiation hardness, high density, and fast access times, simultaneously. Information is stored magnetically in small magnetic elements (micromagnets), allowing unlimited data retention time, unlimited numbers of rewrite cycles, and inherent radiation hardness and SEU immunity, making the MHRAM suitable for ground based as well as spaceflight applications. The MHRAM device design is not affected by areal property fluctuations in the micromagnet, so that high operating margins and high yield can be achieved in large-scale IC fabrication. The MHRAM has short access times (<100 nsec). Write access time is short because on-chip transistors are used to gate current quickly, and magnetization reversal in the micromagnet can occur in a few nanoseconds. Read access time is short because the high electron mobility Hall sensor (InAs or InSb) produces a large signal voltage in response to the fringing magnetic field from the micromagnet. High storage density is achieved since a unit cell consists only of two transistors and one micromagnet-Hall effect (M-H) element. By comparison, a DRAM unit cell has one transistor and one capacitor, and an SRAM unit cell has six transistors.

INTRODUCTION

Ever increasing data processing requirements demand faster and denser random access memory (RAM) to keep pace with improved CPU speed and throughput. Nonvolatility becomes an important factor in many applications where reliability, fault tolerance and fault recoverability, and low power consumption are necessary. Semiconductor memories such as dynamic RAM and static RAM (DRAM and SRAM) have very fast access times, but are volatile, and batteries which could provide backup power are considered risky, unreliable, and consume mass and volume[1]. EEPROMs are non-volatile, but have very long write times (on the order of msec), limited endurance (10^4 to 10^6 write cycles), and require compromises between refresh needs and radiation tolerance. Shadow SRAMs, which have a regular SRAM cell and an EEPROM cell in each memory cell, are costly and still suffer from the endurance problem of the EEPROM. Ferroelectric RAM (FRAM) offers short read and write access times, but the data retention (nonvolatility) and the longevity of the ferroelectric material (reliability) are in question. The magneto-resistive random access memory (MRAM[2]) is non-volatile and has no problem with longevity, but has long read access times (on the order of microseconds).

 PRECEDING PAGE BLANK NOT FILMED

No existing nonvolatile RAM technology satisfies each of the needed data storage requirements. A summary of the performance of these technologies, including the 256 Kbit and 1 Mbit versions of the proposed MHRAM[3], is given in Table 1. It is seen in Table 1 that MRAM and MHRAM come the closest to being the most reliable and truly nonvolatile memories. The proposed MHRAM uses a novel scheme to achieve short read access time while retaining the other merits of the MRAM. The MHRAM therefore has the potential to replace SRAMs and all types of ROMs, including PROMs, EPROMs, UVPROMs and EEPROMs.

POTENTIAL APPLICATIONS

Low Cost, High Performance Replacement for EPROMs and EEPROMs

To write to EPROMs and EEPROMs, a voltage higher than 5 V is usually needed. Special circuit and programming sequence are needed for the write function. The write operation may take from milliseconds to seconds. During these times, the memory content can not be read. Therefore, the write operation is usually done with operator intervention. The reprogramming cost greatly outweighs the cost of the chip. The EPROM and EEPROM can only be reprogrammed for 10^4 to 10^6 times before permanent damage is done. Thus, the user must minimize and keep track of the number of the write cycles, so that the chip can be replaced periodically. The MHRAM can be treated as a regular SRAM. It can be read and written just like any other main memory, using the same instruction and memory cycle time. There is no need to limit the number of write cycles or replace the chip periodically.

Memory Card

Flash EEPROM is currently proposed for the memory card to be used in the notebook computer. Flash EEPROM has the same drawbacks as the EEPROM, i.e., long write time and limited number of write cycles. MHRAM can be used to implement a fast access time and unlimited write cycle memory card.

Solid State Disk

The currently available solid state disks are implemented using SRAM. If non-volatility is required, a battery is used as the backup power. MHRAM can be used to implement a non-volatile solid state disk, which does not need backup power. Solid state disk is typically used as a cache memory between the main memory and the hard magnetic disk. A truly non-volatile solid state disk would relieve the system requirement to backup the content in the solid state disk to the hard disk during power down.

Main Memory

Using MHRAM to implement the main memory can have a profound effect on the computer system. Since the content of the main memory is not lost after the power is lost, the computer can resume its computation after the power has been reinstalled. There will be no need to use the write-through scheme in managing the memory hierarchy, the more efficient write-back scheme can be used without worrying the loss of data due to power failure. There will be other impacts when such a non-volatile memory is available.

MHRAM CELL STRUCTURE

A high speed, non-volatile random access memory cell can be achieved by a micromagnet-Hall effect (M-H) element whose structure is shown in Fig. 1. The M-H element consists of a ferromagnetic element (called micromagnet), shown as the right-slanted area, and a Hall effect sensor, shown as the shaded area. The non-volatile storage function is realized with a micromagnet having an in-plane, uniaxial anisotropy, and, very importantly, in-plane, bipolar remanent magnetization states. The information stored in the micromagnet is detected by a Hall effect sensor which senses the fringing field from the micromagnet. As shown in Fig. 1, when the magnetization in the micromagnet is pointing to the right, a current flowing from lead 3 to lead 4 in the Hall sensor would produce a Hall voltage across leads 1 and 2, with lead 2 being positive with respect to lead 1. When the magnetization in the micromagnet is reversed to point to the left, the same current in the Hall sensor would produce a Hall voltage with the same magnitude but the lead 1 becomes positive. Such a reversal in polarity between the two leads can be detected easily by a differential sense amplifier.

The micromagnet can be magnetized by a local applied field, whose direction is used to form either a "0" or "1" state. The micromagnet remains in the "0" or "1" state until a switching field is applied to change its state, therefore achieving nonvolatility. The micromagnet is magnetized within a few nanoseconds or less, so the write cycle time is very short. The Hall voltage is produced across leads 1 and 2 within a nanosecond. Since this is a differential signal, the settling time of the sense amplifier is short. Both the read and write access times can be within 100 nsec.

MEMORY ORGANIZATION

One organization for a 2x2 bit MHRAM is shown schematically in Fig. 2. The M-H elements are incorporated in a matrix of gating transistors. Micromagnets are shown as rectangles, and Hall sensors are shown as shaded regions. Consider the cell, MH-21, which occurs at the intersection of the second row and the first column. To read the content of MH-21, signals RS2 (Row Select 2), CS1 (Column Select 1), and Read become high. Transistors Q7, Q9, and Q13 are turned on by RS2, which sends a current through, and produces a Hall voltage at, every Hall sensor in the second row. Each Hall voltage is amplified by a sense amplifier at the corresponding column. However, only transistor Q1 has been turned on (by CS1 AND Read), so that only the signal output from sensor 21 is connected to the final output, V_{out}. Although transistor Q8 is also turned on when the second row is selected, since none of the transistors Q2, Q3, Q10, and Q11 are turned on, current does not flow through Q8. If a single-output memory organization is desired, the number of sense amplifiers can be reduced to one if the selection transistors (such as Q1) are placed at the input side of the sense amplifier.

To write to cell MH-21, signals RS2, CS1, and Write become high. If the bit value to be written is a "1," then transistors Q3, Q8 and Q11 are turned on, and if the bit value is a "0," then transistors Q2, Q8 and Q10 are turned on. The bit value then decides the sense of the current through the conductor over the magnetic element, and therefore the sense of the in-plane magnetization. It is noted that the switching current amplitude can be set to accommodate the micromagnet with the highest switching threshold among the matrix of micromagnets. This is because the switching field is applied only to the selected micromagnet and, unlike the MRAM and core memory, will not have a half-select problem. Therefore the writing process achieves high margin and is immune to fluctuations in the switching threshold value, caused for example by material variations, so that high chip yields can be achieved.

READ-OUT SIGNAL LEVEL

One of the critical memory performance parameters is the read-out signal level. The output voltage of the Hall sensor, V_{out}, is given by:

$$V_{out} = \mu V_x B_z W/L, \qquad (1)$$

where μ is the Hall electron mobility, V_x is the voltage drop across the Hall sensor, and W and L are the width and length of the Hall sensor, respectively. It can be seen that the output voltage is proportional to μ. Using the value of μ=1 m^2/V-s, B_z = 100 Oe (0.01 Wb/m^2), and W/L=1, Eq. 1 becomes

$$V_{out} = 0.01 V_x, \qquad (2)$$

i.e., the output is 10 mV when the voltage across the input is 1 V. If the design goal is to have a 10 mV output voltage, with a 5 V supply, neglecting the voltage drop on the gating transistors, a maximum of 5 Hall sensors can be connected in series.

MATERIALS CONSIDERATIONS

Hall Sensor:

It can be seen from Eq. 1, that the Hall sensor's output voltage is proportional to the electron mobility of the Hall sensor. A high electron mobility material is very desirable. At room temperature, the electron mobility in a single crystal silicon is 0.13 m^2/V-s, which is too low for this application. In a bulk single crystal gallium arsenide (GaAs), μ=0.78 m^2/V-s. After doping, μ=0.5 m^2/V-s. In a bulk single crystal indium arsenide (InAs), μ=3.3 m^2/V-s. In an MBE deposited InAs thin film, μ=1.0 m^2/V-s. In a bulk single crystal indium antimonide (InSb), μ=7.8 m^2/V-s. In thermally evaporated InSb thin films[4-9], mobility varies from 0.03 to 6 m^2/V-s, depending on the deposition and annealing conditions. Due to the high cost and low throughput of the MBE, the evaporation and recrystallization process is clearly more practical for mass production.

It is important to note that a 1% variation in the properties of the Hall sensors or the micromagnets throughout the chip only causes a 1% difference in the output voltage. This feature constitutes a significant improvement over the MRAM[2] in which a 1% variation can induce a 50% or 100% variation in output the signal is measured differentially against a fixed reference. Therefore, the MHRAM is expected to be much more insensitive to process variations which occur when manufacturing a large matrix memory.

Micromagnet:

The criteria for choosing the micromagnet material are that it (1) can be switched relatively easily by a field on the order of 100 Oe to limit on-chip power dissipation and current density, and (2) must have a stable non-zero remanence so that it will remember its magnetization state after the switching field is removed. A larger remanence produces a larger fringing field for the Hall sensor. Permalloy has been shown[10-12] to have a coercivity ranging from 25 to 120 Oe. We have shown that CoPt thin films have coercivity ranging from 100 to 600 Oe.

CONCLUSION

We have described a scheme using the combination of a micromagnet, Hall sensor, and gating transistors, to achieve a high speed, non-volatile random access memory. Such a memory has the potential to replace ROMs, EPROMs, EEPROMs, and SRAMs. We have studied the base materials needed to implement this memory. Work is currently underway to implement an integrated circuit prototype of such a memory.

References:

1. B. Gauthier, "CRAF/Cassini Subsystem Memory Maintenance (SCDT AI#307)", Jet Propulsion Laboratory InterOffice Memorandum, 3133-90-106-BS:bs, July 11, 1990.

2. A. V. Pohm, et al., "The Design of a 1 Mbit Non-Volatile M-R Memory Chip", IEEE Trans. Magn. Vol. 24, p. 3117 (1988).

3. J. C. Wu, H. L. Stadler, and R. R. Katti. "High Speed, Non-Volatile Random Access Memory with Magnetic Storage and Hall Effect Sensor." Jet Propulsion Laboratory, New Technology Report, NPO-17999, November 27, 1989.

4. A. R. Clawson, "Bulk-like InSb films by hot wire zone crystallization", Thin Solid Films, Vol. 12, p. 291 (1972).

5. T. Oi, et al, "Microzone recrystallization of InSb film for Hall effect magnetic heads", Jpn. J. Appl. Phys.,

Vol. 17, p. 407 (1978).

6. N. Kotera, Junji Shigeta, et al, "A low noise InSb thin film Hall element", IEEE Trans. Magn., Vol. 15, p. 1946 (1979).

7. M. Oszwaldowski, et al, "Multiple-pass zone melting of InSb thin films", Thin Solid Films, Vol. 85, p. 319 (1981).

8. J. Goc, M. Oszwaldowski, and H. Szweycer, "Dopping of InSb thin films with elements of groups II and VI", Thin Solid Films, Vol. 142, p. 227 (1986).

9. Masaaki Isai, et al, "Influence of thickness on the galvanomagnetic properties of thin InSb films for highly sensitive magnetoresistance elements", J. Appl. Phys., Vol. 59, p. 2845 (1986).

10. S. J. Hefferman, J. N. Chapman, and S. McVitie, "In-situ Magnetizing Experiments on Small Regular Particles Fabricated by Electron Beam Lithography", J. Magnetism and Magnetic Materials, Vol. 83, p. 223 (1990).

11. S. McVitie and J. N. Chapman, "Magnetic Structure Determination in Small Regularly Shaped Particles", IEEE Trans. Magn., Vol. 24, p. 1778 (1988).

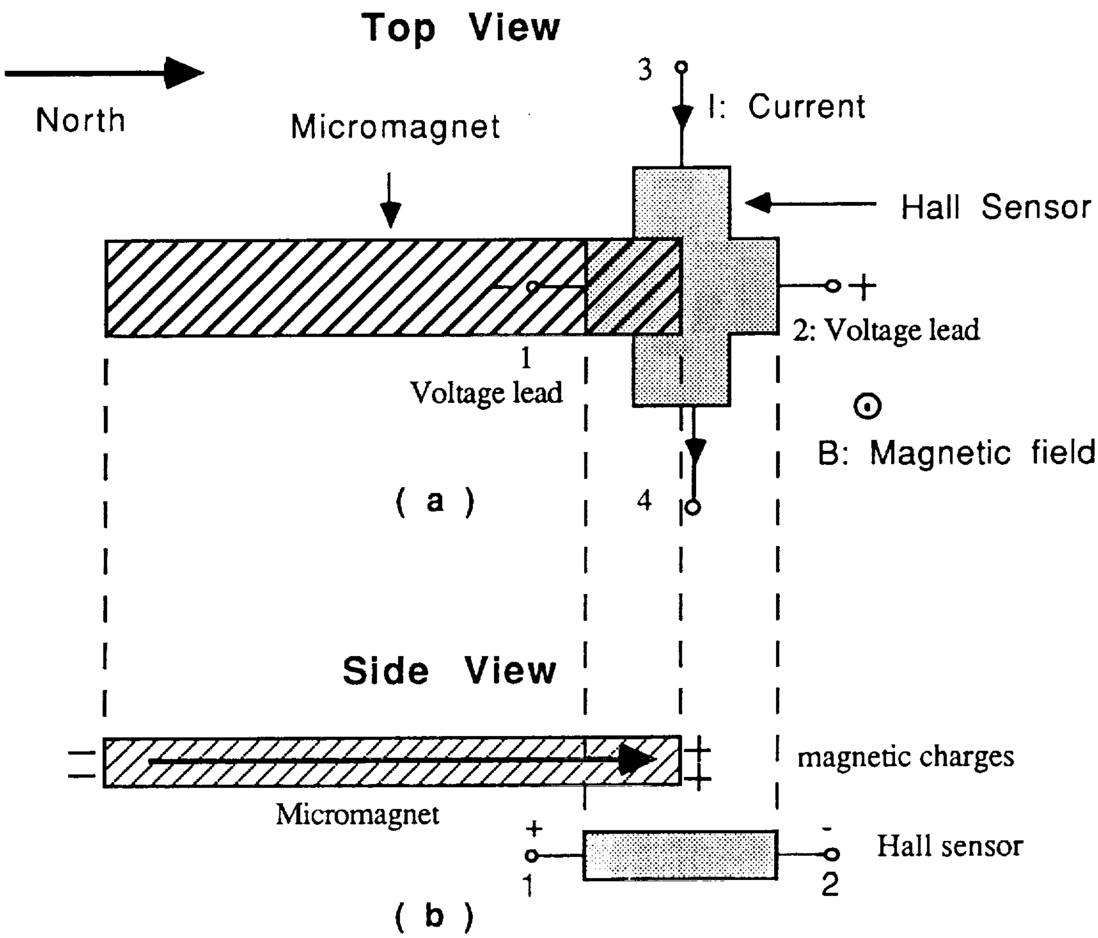
12. J. F. Smyth and S. Schultz, "Hysteresis of Submicron Permalloy Particulate Arrays", J. Appl. Phys., 63 (8), p. 4237, (1988).

Non-Volatile Random Access Memory

Type of Non-volatility	EEPROM	Flash EEPROM	FRAM	MRAM	MHRAM	MHRAM
Sources	Harris Sandia	Intel TI	Krysalis Ramtron	Honeywell	JPL	JPL
Read Cycle Time	200 nsec	150 nsec	200 nsec	1 μsec	100 nsec	100 nsec
Write Cycle Time	10 msec	0.4 msec Erase 3 sec	200 nsec	250 nsec	100 nsec	100 nsec
Capacity (bits)	256 K 1 M	256 K 1 M	16 K	16 K	256 K	1 M
Write Power	200 mW	200 mW	150 mW	150 mW	150 mW	200 mW
Read Power	30 mW	30 mW	150 mW	300 mW	150 mW	200 mW
Endurance	10^4 to 10^5 (write)	10^3 to 10^4 (write)	10^{10} to 10^{12} (access)	Unlimited	Unlimited	Unlimited
Data Retention	hours - 10 yrs	10 yrs	3 yrs	Unlimited	Unlimited	Unlimited
Total Dose Rad(Si)	5E5	N.A.	1E6	1E6	1E6	1E6
Half-select Problem	N.A.	N.A.	N.A.	Yes	Immune	Immune

Table 1

Hall Effect Sensor with In-plane Micromagnet

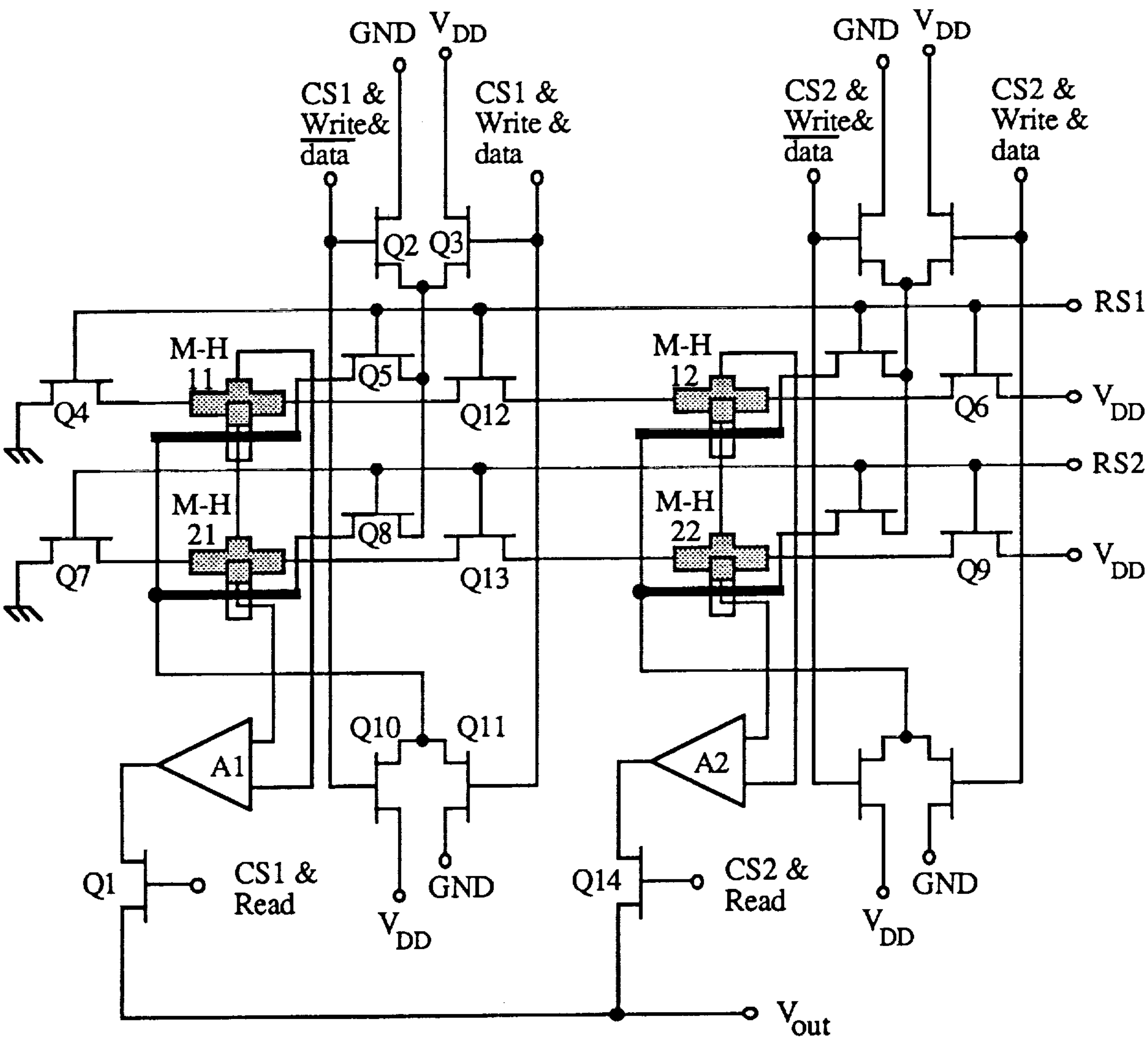


(a) In the absence of the magnet, an n-type Hall sensor, when the current is flowing from lead 3 to lead 4 (east), and an external magnetic field is pointing up, a Hall voltage is developed across leads 1 and 2, with lead 1 being negative.
The polarity of the Hall voltage changes when the magnetic field direction is reversed.

(b) In the presence of a micromagnet magnetized to the right (north), the Hall sensor experiences a downward field, a Hall voltage is developed with lead 1 being positive. When the micromagnet is magnetized south, the polarity of the Hall voltage reverses.

Figure 1

A 2x2 bits MHRAM Organization



CS : Column Select   RS : Row Select   M-H : Magnet-Hall effect element

Two transistors and one M-H element in each unit cell

Figure 2

N92-22441

ANALOG VLSI NEURAL NETWORK INTEGRATED CIRCUITS

F.J. Kub, K.K. Moon and E.A. Just
Code 6813
Naval Research Laboratory
Washington, DC 20375
(202) 767-2534
FAX (202) 767-0546

ABSTRACT

Two analog VLSI vector-matrix multiplier integrated circuit chips have been designed, fabricated and partially tested than can perform both vector-matrix and matrix-matrix multiplication operations at high speeds. The 32x32 vector-matrix multiplier chip and the 128x64 vector-matrix multiplier chip have been designed to perform 300 million and 3 billion multiplications per second, respectively.
An additional circuit that has been developed is a continuous-time adaptive learning circuits. The performance achieved thus far for this circuit is an adaptivity of 28dB at 300KHz and 11dB at 15MHz. This circuit has demonstrated greater than 2 orders-of-magnitude higher frequency of operation than any previous adaptive learning circuit.

INTRODUCTION

Analog VLSI vector-matrix multiplier circuits consist of a two dimensional array of multipliers with the matrix of analog weights stored at the multiplier sites, as shown in Fig. 1. The expression for the vector-matrix multiply operation,

$$V_{yi} = \sum_{j} W_{ij} V_{xj} \qquad (1)$$

where V_{xj} is an input-vector element, W_{ij} is a matrix of weight values, and V_{yi} is an output-vector element. The primary advantages of analog circuits for vector-matrix multiplier operations is that the large two-dimensional matrix of weights (on the order of n^2) are stored at the multiplier sites and do not have to be retrieved from memory as in the digital signal processing case. Only the input vector (on the order of n) has to be retrieved from memory, leading to significant increases in performance.

The vector-matrix multiplier operations is a general function utilized in the vast majority of neural network algorithms and also a large number of conventional signal processing operations. This paper describes fully multiplexed 32x32 and 128x64 programmable analog vector-matrix multiplier circuits. Four-quadrant analog multiplier circuits are used in both chip designs. The weights are X-Y addressed to the multiplier sites and are stored as the difference of analog voltages on two capacitors. (The Naval Research Laboratory holds the basic patent for capacitive weight storage for vector-matrix multiplier circuits [1].) Analog multiplexers are used for the analog input vector, the analog output vector, and the X-Y weight address. A fully differential design has been used throughout the signal path for cancellation of common-mode noise feedthroughs and reduction of offsets. The 32x32 and 128x64 vector-matrix multiplier circuit have been fabricated using a N-well CMOS foundry process. Possible applications of the vector-matrix multiplier circuits are implementing artificial neural network algorithms, implementing large banks of two or three dimensional convolution filters, and performing large area, high speed (1000 frames per second) two dimensional template matching in the Fourier domain.

An additional circuit that has been developed is a continuous-time adaptive learning circuits. The performance achieved thus far for this circuit is an adaptivity of 28dB at 300KHz and 11dB at 15MHz. This circuit has demonstrated over 2 orders-of-magnitude higher frequency of operation than any previous adaptive learning circuit. These circuits can be used as interference cancelers, linear predictors or equalizers. Possible commercial applications areas are smart house wiring over exiting power lines, removal of coherent noise in musical systems, or equalizers for modems or magnetic heads.

32X32 VECTOR-MATRIX MULTIPLIER CHIP

Chip Description

The chip block diagram is shown in Fig. 1. The circuit consists of a two dimensional array of analog multipliers, X-Y address weight decoders, input and output decoders, and current-to-voltage (I-V) converters at the output of each row. In this configuration, the multiplication of a matrix and a vector is performed by capacitively storing analog weights as differential voltages, $\Delta Vw_{ij} = Vw_{ij} - Vwr$, at each multiplier site, and applying a vector of analog inputs as differential voltages, $\Delta V_{xj} = V_{xj} - V_{xr}$, to the column busses. The output currents of the multipliers in a given row are summed on a bus and converted to voltages by the I-V converters to provide the output analog vector elements.

The circuit design approach for fully differential operation is shown in Fig. 1. Both the weight input and weight reference are sampled simultaneously so that feedthroughs from the switches are canceled by the common-mode differencing operation of the four-quadrant analog multipliers. The same procedure is used for the V_x inputs. This technique allows the use of differential D/A converters.

The PMOS Gilbert four-quadrant analog multiplier used in the 32x32 circuit is shown in Fig. 2. The weight values are capacitively stored at the gates of M1 and M2. The difference in current outputs is proportional to the product of two differential voltages, ΔVx and ΔVw. Row and column decoders are used to write the analog voltages to the capacitive weight storage nodes at each of the multiplier sites.

The I-V converters convert the currents on the row busses into voltages. Variable gain circuits are used to optimize the dynamic range. The gain stage is followed by a sample-and-hold circuit that is used to isolate the output from the input so that a new analog vector can be inputted to the vector-matrix multiplier circuit while reading the present output analog vector elements.

Circuit Characteristics

The double-capacitor storage arrangement, shown in Fig. 2, tends to cancel the effects of leakage currents at the capacitor storage sites, thereby significantly improving the weight retention. Measurements of the weight retention shown in Fig. 3 show a factor of 50 improvement for the double- capacitor configuration over the single-capacitor configuration. The bottom curve for the double-capacitor approach shows a 3mV change in the effective weight over a 10mS period at 90C. This change would provide better than 1 percent accuracy for the stored weights. It is necessary that the capacitively stored weights be refreshed. A refresh period of 10ms is reasonable for most applications. Measurements have shown that individual Gilbert multipliers have a total harmonic distortion less than 1.5 percent [2,3].

The cell size for the multiplier shown in Fig. 2 using 2μm design rules is 58μm x 60μm. Figure 4 shows the operation of the fully multiplexed 32x32 vector-matrix multiplier for the case of "1" and "0" weights loaded on alternating rows. An alternating sequence of "1s" and "0s" is observed at the output as expected. A low noise analog board is currently being implemented which will allow further characterization of the circuit's dynamic range.

128X64 VECTOR-MATRIX MULTIPLIER CHIP DESIGN

A schematic of the overall architecture of the 128x64 vector-matrix multiplier circuit is shown in Fig. 5. Four 1-to-32 analog multiplexers are used for inputting the 128 element analog input vector and two 1-to-32 analog mutiplexers are used for outputting the analog vector. The input decoders and weight decoders control transmission switches as shown in Fig. 1. Each of the input and weight decoders is a five-bit random-address decoder with an enable. The five-bit input, output, and weight decoders can address the eight 32x32 multiplier array blocks in parallel when all decoders are enabled. Alternately, if only one of the input, output, or weight decoders is enabled, one individual analog input, one individual analog output, or one individual analog weight can be addressed. The currents from 128 multipliers on each row are summed to produce the analog row output.

The row decoder is used to select 1 of 64 rows to write weights to. It is expected that the input and output multiplexer will operate at approximately 10MHz, thus requiring approximately 3μs to load the input vector. The projected performance for this circuit is approximately 3 billion connections per second.

A feature incorporated into the 128x64 vector-matrix multiplier is a power-saving mode of operation [4] which will likely reduce the power dissipation of large vector-matrix multiplier array by greater than an order-of-magnitude. It can be expected for large vector-matrix multiplier arrays that the power dissipation of each multiplier will be approximately 160μW. Thus, for arrays with 10^4 to 10^5 multipliers, the power dissipation can be in the range of 1.5W to 15W. The power saving mode in the 128x64 vector-matrix multiplier is achieved by turning on the MOSFET triode-mode I-V converter transistors only during the time that the sample-and-hold circuits are turned on. Simulations indicate that the summing bus turn-on time is less than 10ns and the turn-off time is about 30ns. These results indicate that the time period that the triode MOSFETs must be on can be < 100ns. Since the time to load the input vector will be typically 1 to 3μs, this power-saving mode will likely provide greater than an order-of-magnitude reduction in power dissipation.

A photomicrograph of the 128x64 programmable analog vector-matrix multiplier circuit fabricated using 2μm N-well CMOS foundry process is shown in Fig. 6. The chip size is 6.5mm x 6.5mm and the differential-pair multiplier cell size is 31μm x 60μm.

MATRIX-MATRIX MULTIPLY AND HIGH SPEED TEMPLATE MATCHING

The 32x32 and 128x64 circuits described above can also implement a matrix-matrix multiplication operations. The approach is to load one of the matrices into the two-dimensional array of multipliers and then to sequentially input the second matrix a row at a time to the input of the chip. The product of the matrix-matrix multiply operation appears a column at a time at the output of the chip. The multiplication rate performance is the same as that for the vector-matrix multiplier case as long as the first matrix is not reloaded. The performance degrades approximately a factor of two if the first matrix is reload for each matrix-matrix multiply operation. Two of the 128x64 circuits can be used to implement a 128x128 size matrix-matrix multiply operation. The performance rate for the 128x128 size matrix-matrix multiply operation is approximately 6 billion multiplications per second (assuming the matrix stored in the chip is reloaded infrequently).

Two dimensional template matching in the Fourier domain consists simply of a multiplication of a reference template matrix by an image template matrix. Fourier transforms of two dimensional images are readily performed by using conventional digital Fast Fourier Transform (FFT) circuits. Two one dimensional FFT transform operations plus a reformatting operation are necessary to implement a two dimensional Fourier transform. FFT circuits are now available commercially that will implement a 1024 point transform in approximately 100 microseconds. The envisioned operations of the template matcher is that a two dimensional image would be loaded in the vector-matrix multiplier circuit and reference templates would be applied at a high rate to determine the best match. For the 128x128 circuit, the time required to input a 128x128 reference template is approximately 400 microseconds. Thus, greater than 1000 template match comparisons per second are possible. Alternately, the reference template could be loaded in the circuit and two dimensional images applied at a high rate to determine the best match.

CONTINUOUS-TIME ADAPTIVE LEARNING CIRCUITS

A new approach for high frequency adaptive learning circuits using a continuous-time circuit, shown in Fig. 7, to implement the least mean square learning algorithm has been developed. Previous analog adaptive learning circuits have utilized either CCD sampled data circuits or switched-capacitor circuits. Previously, the highest performance achieved using analog circuits was for a four channel switched-capacitor circuit operating at 40KHz [5]. The advantages of the continuous-time approach are the potential for a large number of adaptive taps (>200) and the potential for a high frequency of operation (>50MHz).

The performance achieved thus far is an adaptivity of 28dB at 300KHz and adaptivity of 11dB at 15MHz (Fig. 8) in the linear predictive arrangement. Also, a notch filter with a 10KHz half-width and an isolation of 30dB (Fig. 9) was achieved for an interference canceler arrangement.

This adaptive filter circuit has been fabricated using 2μm CMOS foundry technology. The cell size for the learning circuitry at each tap is 43μm x 1150μm. Thus, an adaptive learning circuit with a large number of taps can be achieved in a typical integrated circuit chip size.

CONCLUSIONS

Fully multiplexed vector-matrix multiplier circuits using capacitive weight storage with a standard N-well CMOS technology have been described. Experimental results were presented for the operation of the Gilbert multiplier, and the 32x32 vector-matrix multiplier circuit. A method of reducing power dissipation by greater than an order-of magnitude was described. A fully multiplexed 128x64 analog vector-matrix multiplier circuit was implemented in an area of 6.5mm x 6.5mm using 2μm CMOS foundry design rules. A method to perform high frame rate (>1000 frames per second), large size (128x128) two-dimensional template matching was described.

An additional circuit type that has been developed is a continuous-time adaptive learning circuit. This circuit has thus far demonstrated high levels of adaptivity and greater than two orders-of-magnitude improvement in the frequency of operation over any previous analog learning circuit. Circuit designs to achieve higher levels of performance are being implemented.

ACKNOWLEDGMENT

The authors would also like to acknowledge the support of the Office of Naval Research and the Office of Naval Technology for this work.

REFERENCES

[1] Patent 4,931,674, F.J. Kub, I.A. Mack, and K.K. Moon, Programmable analog voltage multiplier circuit means, Issue Date: 5 June 1990.

[2] F.J. Kub, K.K. Moon, I.A. Mack and F.M. Long, "Programmable analog vector-matrix multiplier," IEEE Journal of Solid State Circuits, Vol. SC-25, pp. 207-214, February 1990.

[3] K.K. Moon, F.J. Kub and I.A. Mack, "Random address 32x32 programmable analog vector-matrix multiplier for artificial neural networks," Proceedings of the 1990 Custom Integrated Circuit Conference, May 13-16, Boston, Mass., pp. 26.7.1-26.7.4.

[4] F.J. Kub, K.K. Moon and J.A Modolo, "Analog Programmable Chips for Implementing ANNs using Capacitive Weight Storage," To be published in Proceedings of International Joint Conference Neural Networks-91-Seattle, July 8-12, 1991.

[5] J. Fichtel, J.H. Hoticka, and P. Sieber, "An analog adaptive filter in BICMOS Technology," Proc. of ISSCC Conf., San Francisco, 1990.

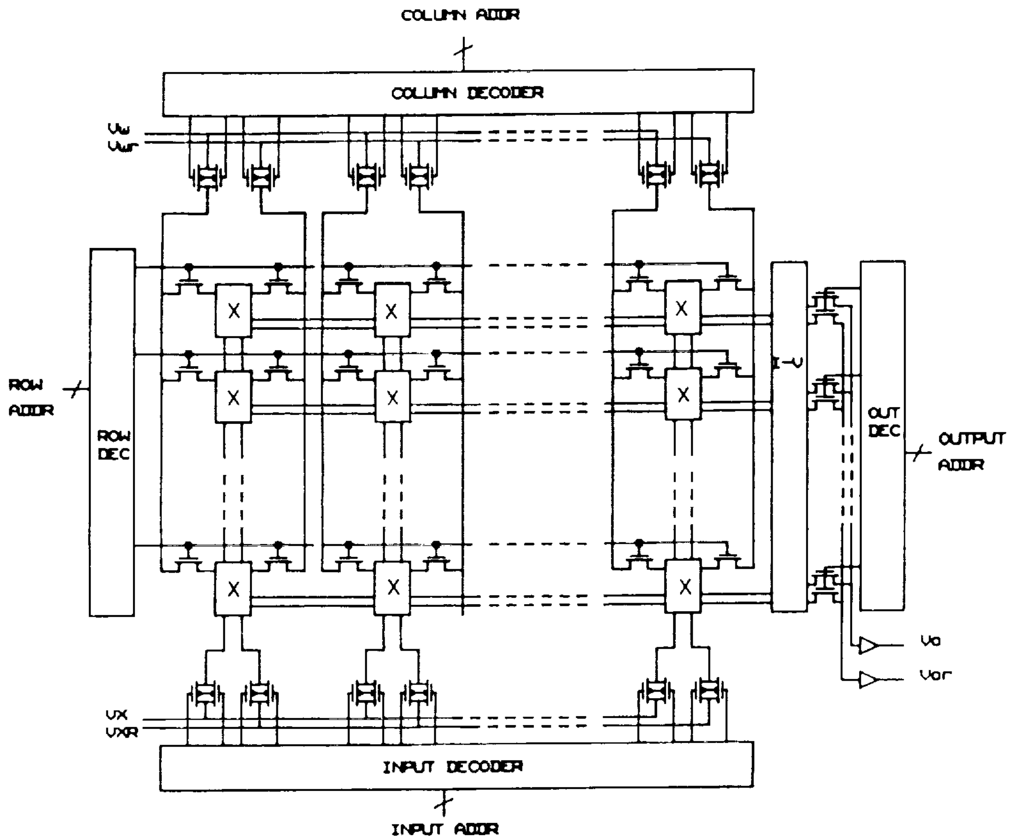


Fig. 1. Vector-matrix multiplier block diagram.

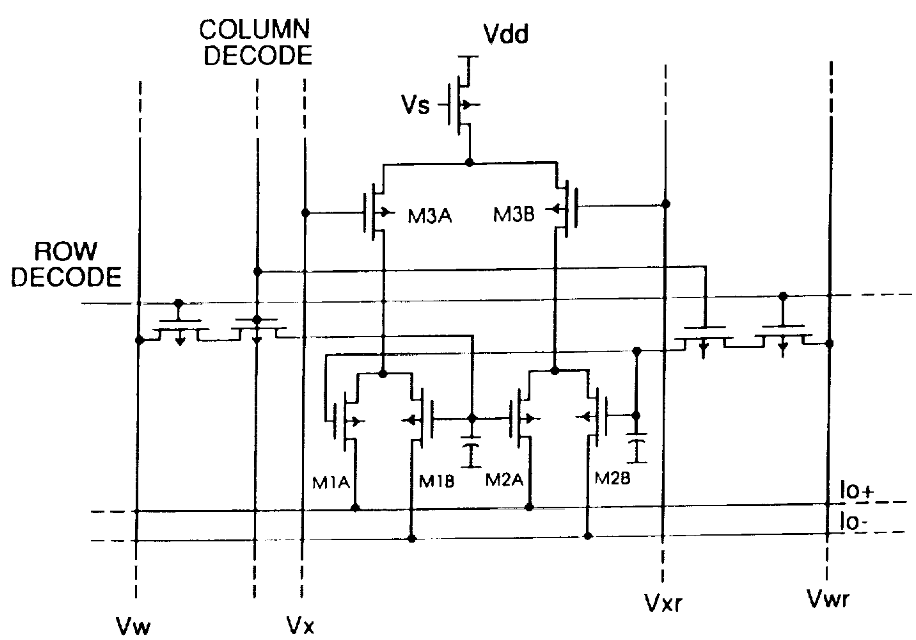


Fig. 2. Gilbert multiplier cell with capacitive weight storage.

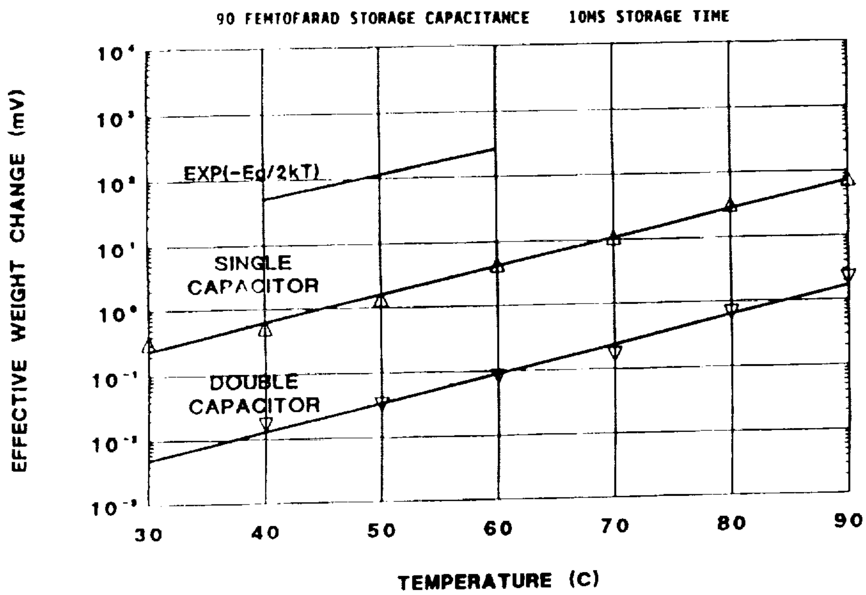


Fig. 3. Weight retention versus temperature.

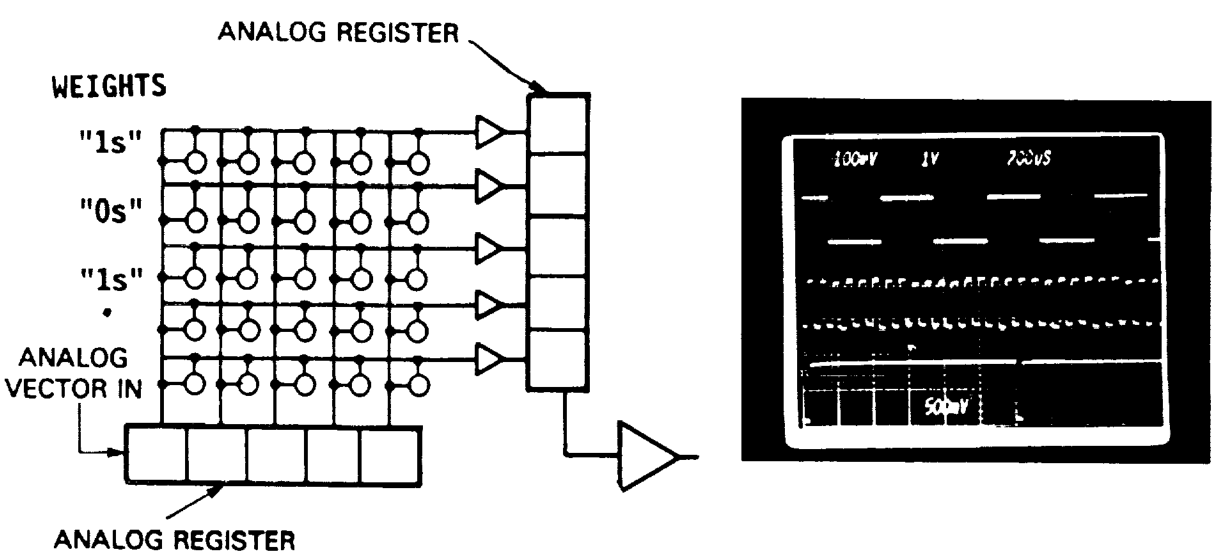


Fig. 4. Output of 32x32 vector-matrix multiplier with "1s" and "0s" loaded on alternating rows.

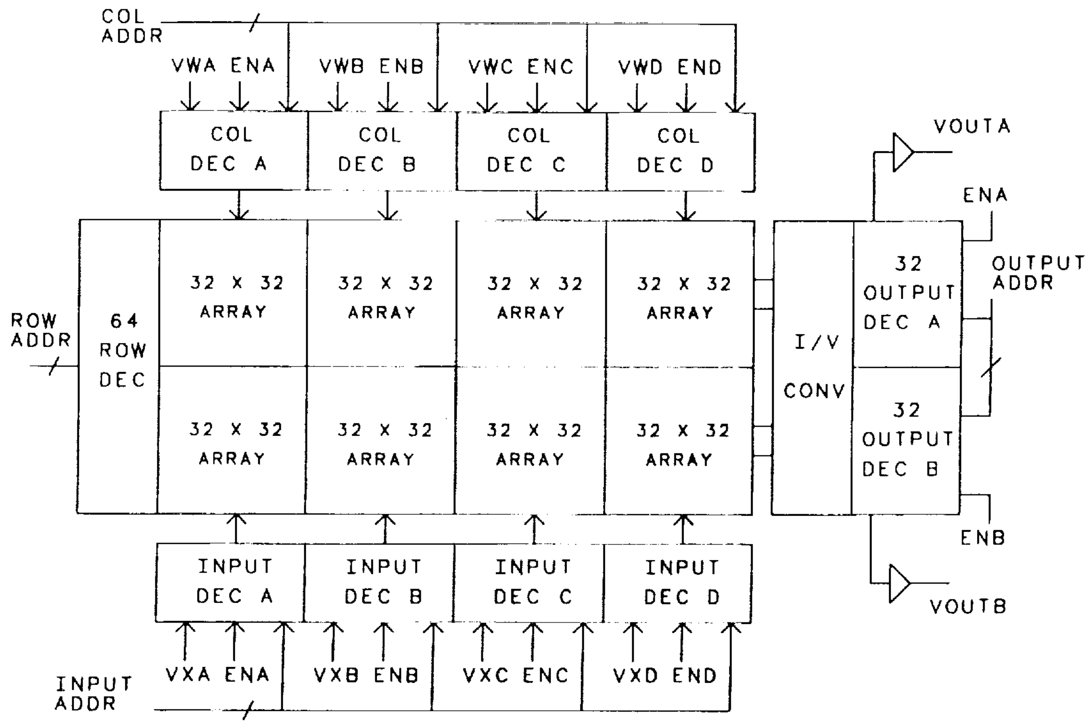


Fig. 5. Architecture of 128x64 vector-matrix multiplier.

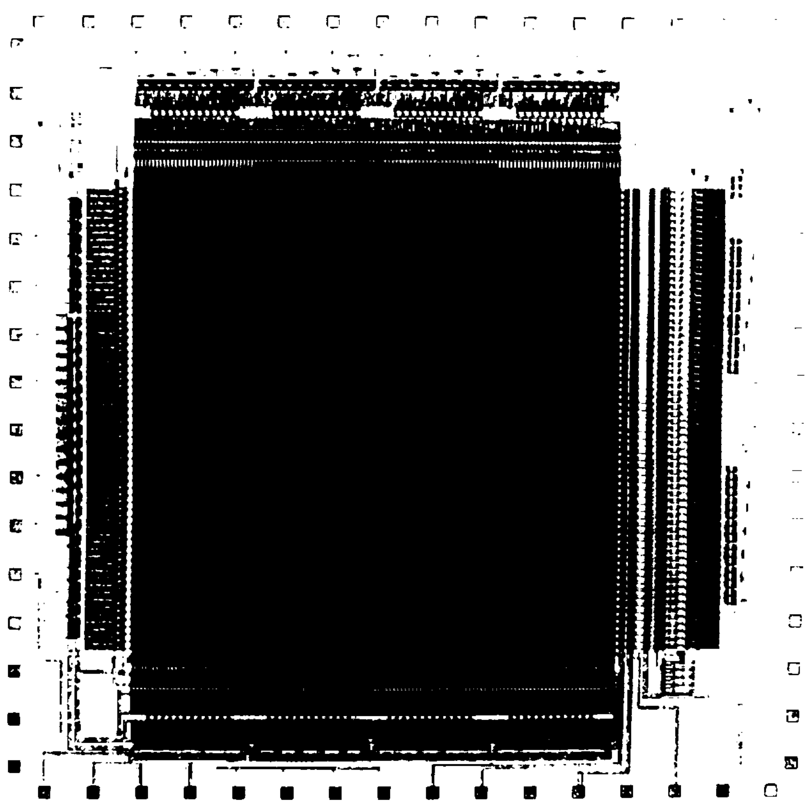


Fig. 6. Photomicrograph of 128x64 vector-matrix multiplier chip.

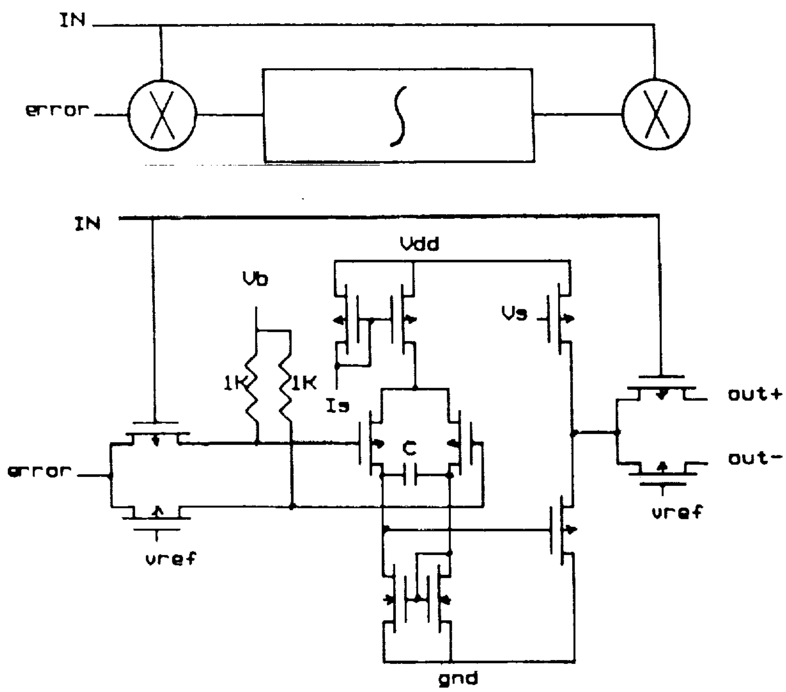
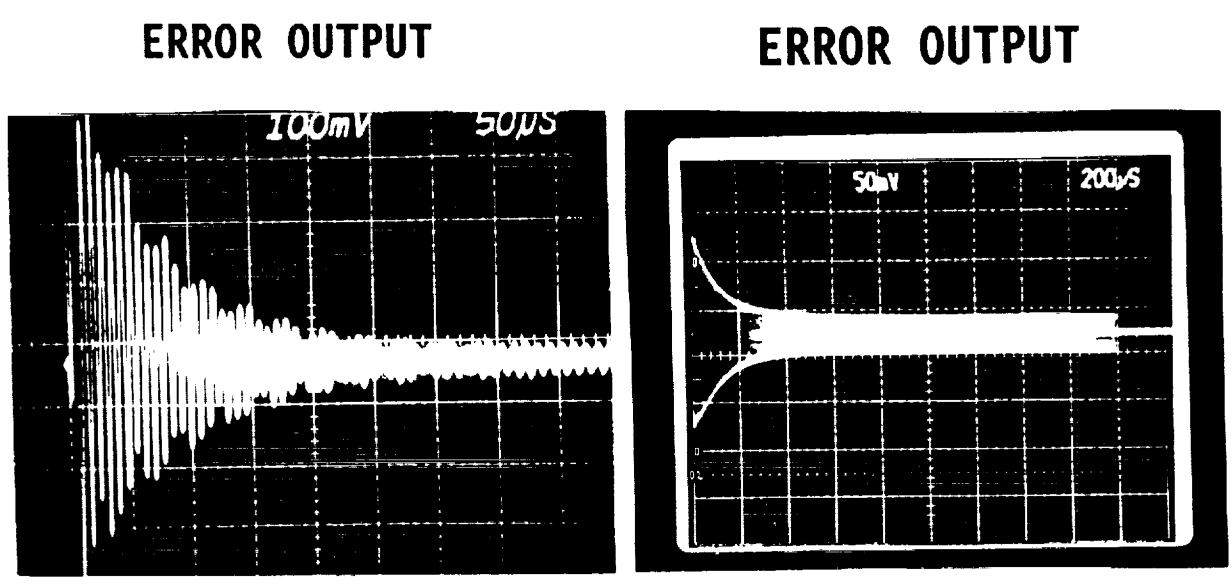
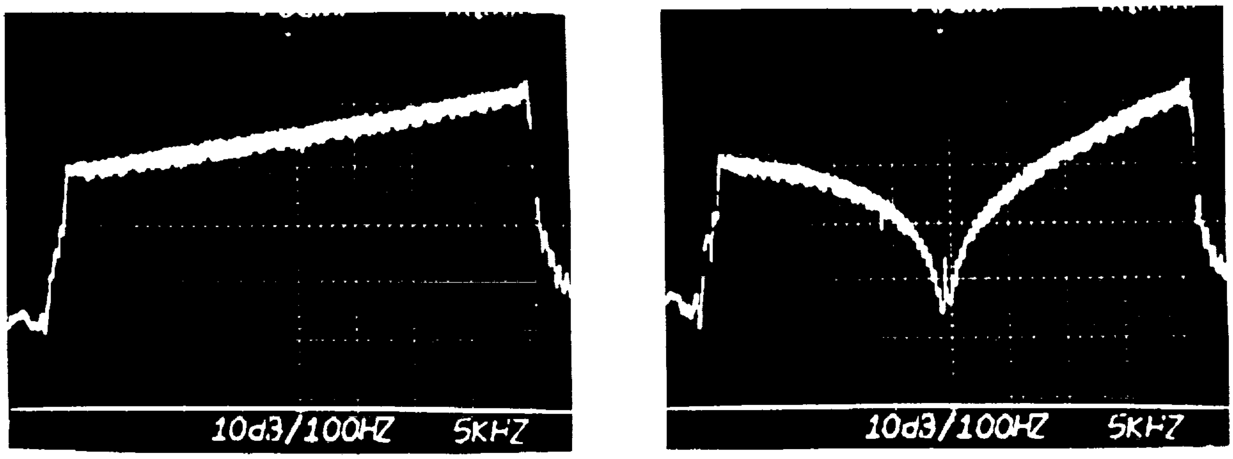


Fig. 7. Continuous-time least mean square weight learning circuit.



(100KHz) (15MHz)

Fig. 8. Experimental results showing error output for linear predictive arrangement.



SIGNAL IN **NOTCH FILTER OUTPUT**

Fig 9. Experimental results showing adaptive filter operated as a notch filter (80kHz).

N92-22442

Monolithic Microwave Integrated Circuit Water Vapor Radiometer

L.M. Sukamto, T.W. Cooley, M.A. Janssen, G.S. Parks
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109

Abstract

A proof-of-concept Monolithic Microwave Integrated Circuit (MMIC) Water Vapor Radiometer (WVR) is under development at Jet Propulsion Laboratory. WVRs are used to remotely sense water vapor and cloud liquid water in the atmosphere and are valuable for meteorological applications as well as the determination of signal path delays due to water vapor in the atmosphere. The high cost and large size of existing WVR instruments motivate the development of miniature MMIC WVRs, which have great potential for low cost mass production. The miniaturization of WVR components allows large-scale deployment of WVRs for Earth-environment and meteorological applications. Small WVRs can also result in improved thermal stability, resulting in improved calibration stability. This paper describes the design and fabrication of a 31.4 GHz MMIC radiometer as one channel of a thermally stable WVR, to assess the MMIC technology feasibility.

Introduction

A WVR is "... a device for measuring sky brightness temperature at two frequencies on and near the emission line at 22.2 GHz."[1] A typical WVR consists of two independent radiometers, or channels, tuned at 20.7 GHz and 31.4 GHz. The development effort described here concentrated on the design and fabrication of the 31.4 GHz channel alone, with the idea that the second (e.g. 20.7 GHz) channel could be constructed in the same way.

A primary motivation for this research is to build a thermally stable WVR through the minimization of the WVR components. Improved WVR thermal stability should lead to increased system calibration stability. Dual channel radiometers, at 20.7 GHz and 31.4 GHz, sense sky brightness temperatures which respond to both water vapor and liquid water. The 20.7 GHz channel is more affected by water vapor, while the 31.4 GHz channel by liquid water. Two independent and simultaneous measurements of sky brightness temperatures at these frequencies allow the extractions of water vapor and liquid water content in the atmosphere.[2] Measurement accuracy can be increased with improved system calibration stability.

A thermally stable WVR is therefore desirable. By miniaturizing the WVR instrument, thermal stability can be improved.

System Design

The system design approach was to integrate MMIC chips from commercial and research foundries onto carriers built at JPL, and also to develop a modular package design which allows individual modules to be tested. The WVR receives a Radio Frequency (RF) signal with a 400 MHz bandwidth centered at 31.4 GHz, amplifies the signal, downconverts it to an Intermediate Frequency (IF) of 9.4 GHz, detects the IF signal, then processes it to give an output frequency between 0-100 KHz. This output frequency is proportional to the measured sky temperature.

A system block diagram is shown in Figure 1. The radiometer MMIC assembly consists of a noise source module, a Low Noise Amplifier (LNA) module, a mixer module, an IF amplifier and bandpass filter module, and a detector/voltage-to-frequency converter module. The noise source module is used for system calibration. The LNA module amplifies the RF signal and determines the system noise figure. Frequency downconversion is done through the mixer. The IF amplifier and bandpass filter module serves as a gain block and determines the system bandwidth. The detector/voltage-to-frequency converter module detects the IF signal then converts it to a frequency pulse between 0-100 KHz which is proportional to the amplitude of the IF signal. A photograph of the radiometer MMIC assembly is shown in Figure 2 and the individual MMIC modules are shown in Figure 3. These MMIC modules along with a Dielectric Resonator Oscillator (DRO) for the mixer Local Oscillator (LO) supply and an antenna will be integrated onto a heat sink carrier, as illustrated in Figure 4. The LNAs and IF amplifier chips were procured from commercial foundries (Varian and Pacific Monolithics), and the mixer was developed and fabricated at Honeywell. The DRO was purchased from Varian.

Among the challenges of this effort are the package design, the electrical performance characterization of the MMIC chips, and the fabrication and assembly of the modules. MMIC modular testing, minimization of size and thermal analysis are major considerations in the package design. The MMIC chips were mounted on their module carriers for testing. Module carriers were machined from Molybdenum. The rest of the WVR package was made from brass. Package fabrication and module assembly were done at JPL.

Module testing is currently in progress. Preliminary tests on the LNA module, mixer, IF amplifier module and the detector/ voltage-to-frequency module have been performed, and system integration and test will follow.

Applications

The potential commercial applications of WVRs are promising. The National Oceanic and Atmospheric Administration's (NOAA) ground based dual-wavelength WVRs have shown the ability to monitor aircraft icing conditions by measuring supercooled liquid water in clouds. Such radiometers, if they could be made small and at low cost, could be deployed at small and medium sized airports. NOAA has a related application in which measurements of precipitable water vapor can be used by numerical weather prediction models; such data could be used to improve the calibration of polar orbiting and geostationary satellites. The Department of Energy (DOE) is investigating the implementation of numerous WVRs at several global study sites over the next two years for their Atmospheric Radiation Monitoring (ARM) project. The ARM project seeks to improve the capability of general circulation models so that global effects, such as global warming, can be predicted.

Conclusions

This proof-of-concept 31.4 GHz MMIC WVR demonstrates the technological feasibility of a miniature radiometer instrument. The cost of large-scale production of this MMIC WVR is potentially low, which is essential in commercial applications.

Acknowledgements

The research described in this paper was performed by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. The authors wish to thank Dr. Ed Westwater for his contributions on NOAA radiometer applications, and to acknowledge G. Boreham, S. Chavez, C. Cruzan and C. Jones for their contributions in fabrication and assembly.

References

[1] G.M. Resch, *et al.* "Description and Overview of an Instrument Designed to Measure Line-of Sight Delay Due to Water Vapor" *TDA Progress Report 42-72*, October-December 1982, pp. 4, 5.

[2] D.C. Hogg, *et al.* "Measurement of Excess Radio Transmission Length on Earth-space Paths" *Astronomy and Astrophysics*, 95, 1981, pp. 304, 305.

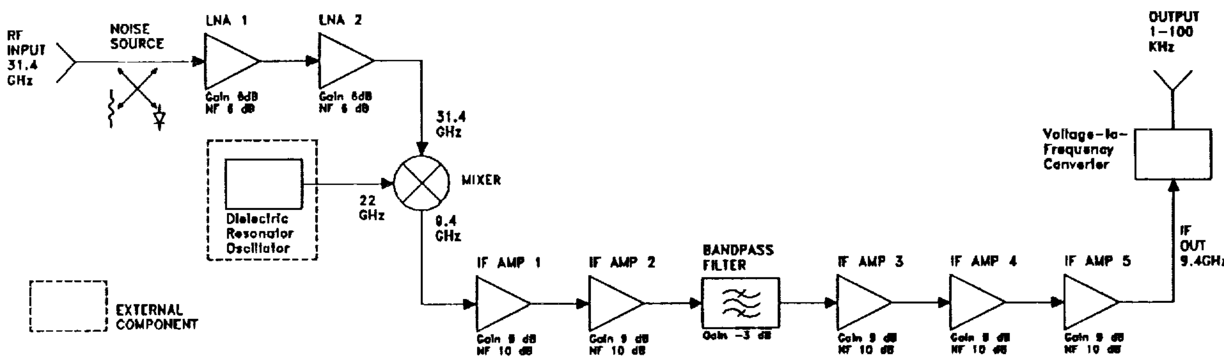


Figure 1. System Block Diagram

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

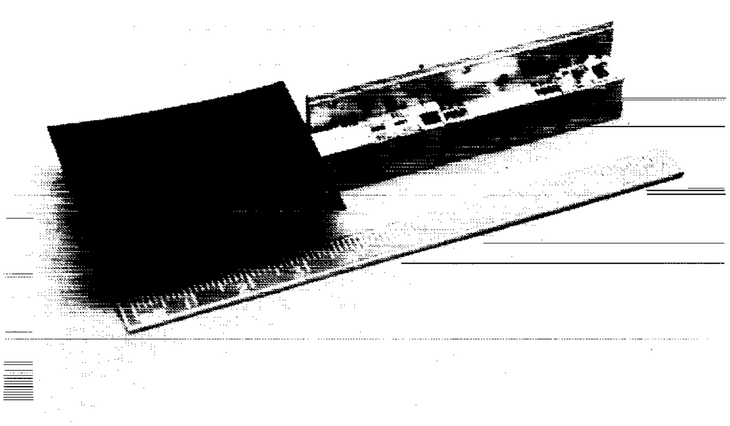
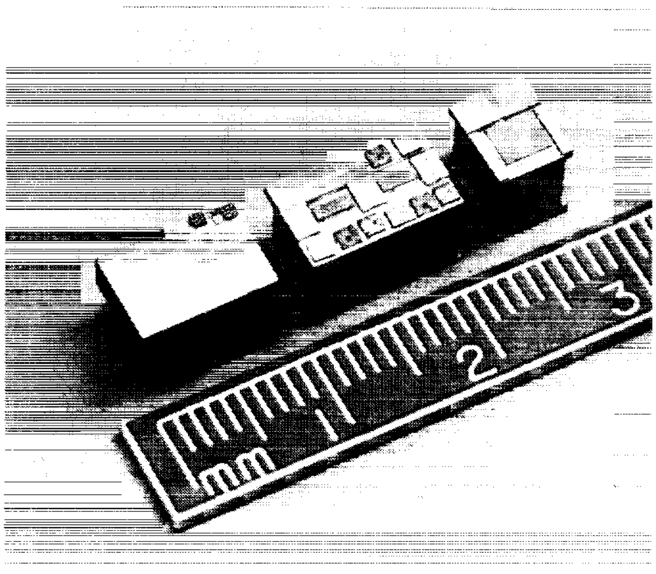
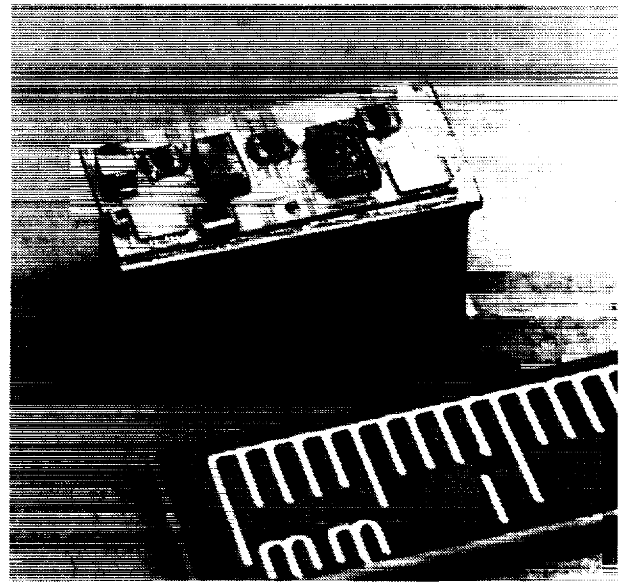


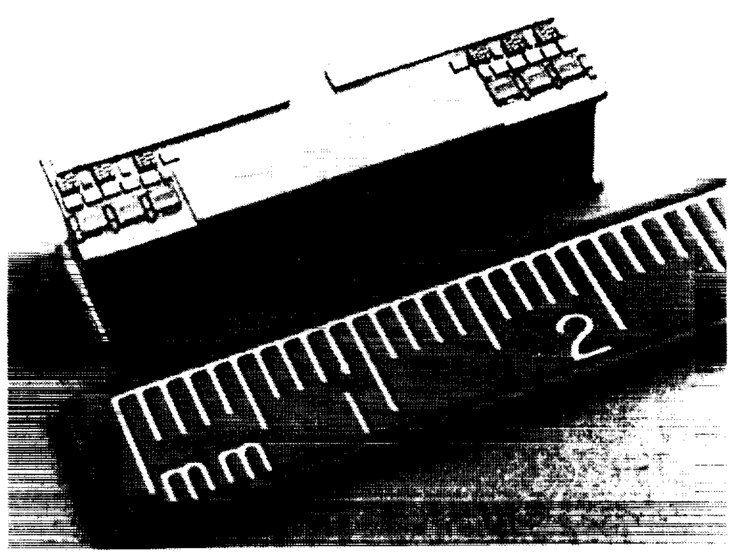
Figure 2. MMIC Radiometer Assembly



Noise Source,Low Noise Amplifier and Mixer Module



Detector/Voltage-to-Frequency Converter Module



IF Amplifier and Bandpass Filter Module
Figure 3. MMIC Modules

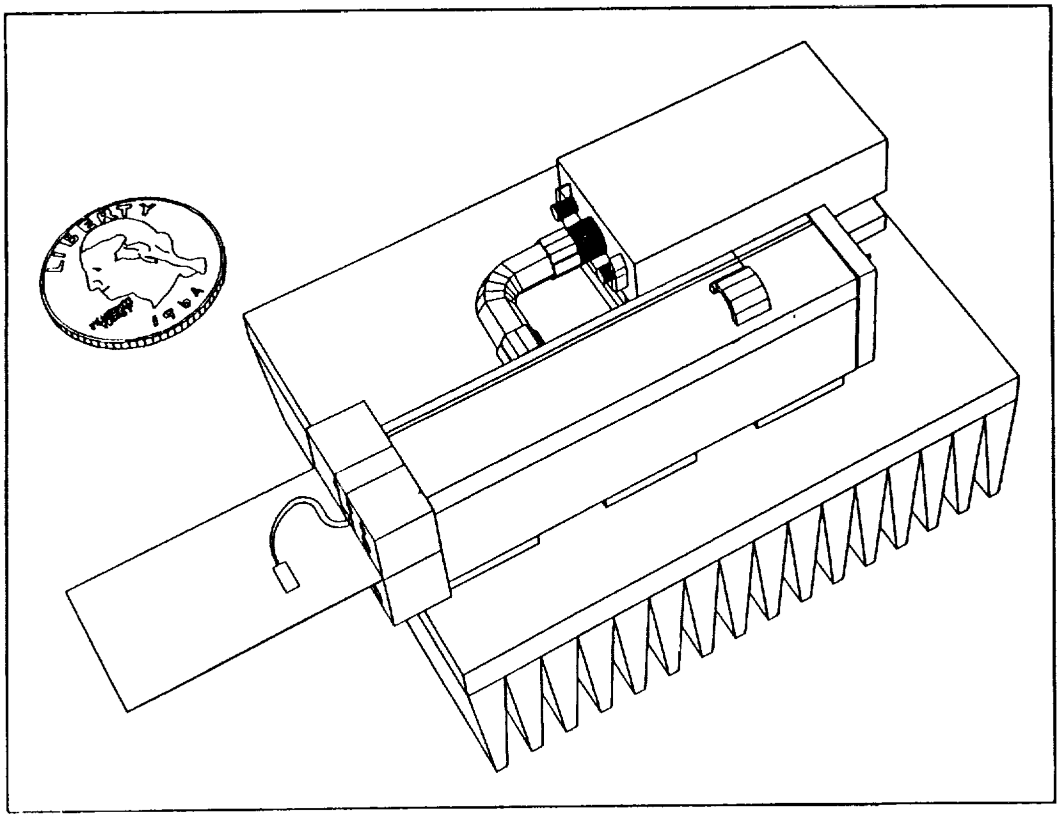


Figure 4. MMIC Radiometer Integrated System Layout

N92-22443

A NOVEL NONCONTACTING WAVEGUIDE BACKSHORT FOR MILLIMETER AND SUBMILLIMETER WAVE FREQUENCIES

W. R. McGrath
Center for Space Microelectronics Technology
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109

ABSTRACT

A new noncontacting waveguide backshort has been developed for millimeter and submillimeter wave frequencies. It employs a metallic bar with rectangular or circular holes. The size and spacing of the holes are adjusted to provide a periodic variation of the guide impedance on the correct length scale to give a large reflection of rf power. This design is mechanically rugged and can be easily fabricated for frequencies from 1 GHz to 1000 GHz. This design is particularly useful for submillimeter wave frequencies above 300 GHz where conventional backshorts are difficult to fabricate. Model experiments have been performed at 4-6 GHz to optimize the design. Values of reflected power greater than 95% over a 30% bandwidth have been achieved. The design has been scaled to WR-10 band (75-110 GHz) with comparably good results.

INTRODUCTION

Waveguides are used in a wide variety of applications covering a frequency range from 1 GHz to over 600 GHz. These applications include radar, communications systems, microwave test equipment, and remote-sensing radiometers for atmospheric and astrophysical studies. Components made from waveguides include transmission lines, directional couplers, phase shifters, antennas, and heterodyne mixers, to name a few. In addition to the many commercial applications of waveguides, NASA needs such components in radiometers operating up to 1200 GHz for future space missions, and the Department of Defense is interested in submillimeter wave communications systems for frequencies near 1000 GHz.

One of the most frequent uses of waveguide is as a variable length transmission line. These lines are used as tuning elements in more complex circuits. Such a line is formed by a movable short circuit or "backshort" in the waveguide. A conventional approach is to use a contacting backshort where a springy metallic material, such as beryllium copper, makes DC contact with the broadwalls of the waveguide. The contacting area is critical, however, and must make good contact to produce an acceptable short circuit. These backshorts are excellent in that they provide a good short circuit over the entire waveguide band. However, the contacting areas can degrade from sliding friction and wear. It is also extremely difficult to get a uniform contact at frequencies above 300 GHz where the waveguide dimensions become 0.5 mm × 0.25 mm for the 300-600 GHz band.

An alternative solution is the noncontacting backshort shown in Fig. 1. A thin mylar insulator prevents contact and allows the backshort to slide smoothly. In order to produce an rf short circuit and, hence, a large reflection, this backshort has a series of high- and low-impedance sections which are approximately $\lambda g/4$ in length where λg is the guide wavelength. The rf impedance of this design is given approximately by

$$Z_{rf} = \left(\frac{Z_{low}}{Z_{high}}\right)^n Z_{low} \tag{1}$$

where Z_{low} is the guide impedance of the thick (low-impedance) section; Z_{high} is the impedance of the thin (high-impedance) section; and n is the number of sections. Values of $Z_{rf} < 1$ ohm are theoretically possible which provides a good short circuit. However, beginning near 100 GHz, the thin high-impedance section become difficult to readily fabricate, and in the 300-600 GHz band, these sections become too thin to fabricate. The backshort is no longer strong enough to slide snugly in the waveguide.

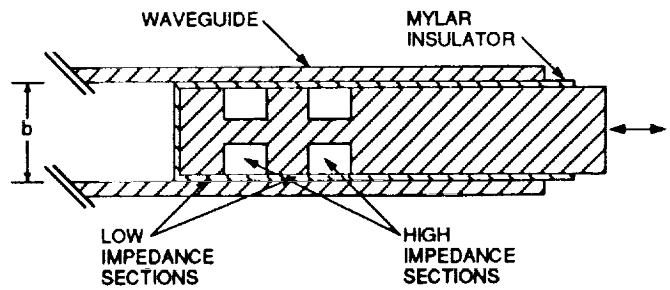


Figure 1. **Cross sectional view of a conventional noncontacting backshort. "b" is the waveguide height.**

New Backshort Design

A new noncontacting backshort has been developed and is shown in Fig. 2. In order to obtain a large reflection, a noncontacting backshort must provide a periodic variation of guide impedance on the correct length scale. This is accomplished in the new design by either rectangular or circular holes with the proper dimensions and spacing cut into a metallic bar. This bar is dimensioned to form a snug fit in the waveguide with a mylar insulator along the broadwalls. The holes replace the thin-metal, high-impedance sections in the conventional design shown in Fig. 1. Since the holes extend completely through the bar, this yields a

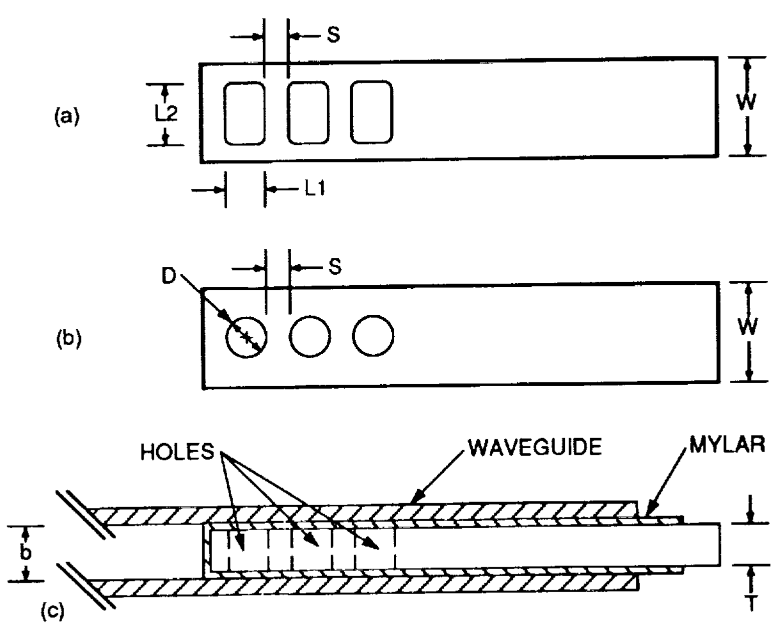


Figure 2. **New noncontacting backshort design. (a) A metallic bar of width W and thickness T with rectangular holes cut near one end. The hole length L_1 and separation S are important in determining the rf properties. "S" is also the distance from the end of the bar to the edge of the first hole. (b) A similar backshort design using round holes. (c) Cross sectional view of the new backshort design in the waveguide. A thin mylar insulator allows the backshort to slide smoothly.**

higher impedance than the corresponding sections in the conventional design. Thus, the high-to-low impedance ratio is larger in the new design. In addition, the electromagnetic fields and power are concentrated near the central axis of the waveguide, so the holes are effective in producing correlated reflections leading to an overall large reflection of rf power. The new design is also easy to fabricate and can be used at any waveguide frequency between 1 GHz and 1000 GHz. For very high frequencies, above 300 GHz, the metallic bar is a piece of shim stock polished to the correct thickness. The holes can be formed by drilling, punching, laser machining, or can be etched using common lithography techniques.

Measurement Techniques

The backshort design was optimized by testing the performance in WR-187 band waveguide (3.16 GHz - 6.32 GHz). The waveguide dimensions are 47.5 mm x 22.1 mm (1.87 in x 0.87 in). The magnitude and phase of the reflection coefficient were measured with an HP 8510B Vector Network Analyzer. A commercially available coaxial-to-waveguide transition was used to connect the waveguide to the network analyzer. This measurement system was calibrated using two offset contacting shorts set at $\lambda g/8$ and $3\lambda g/8$ in the waveguide and a sliding waveguide load. Subsequent verification using a contacting short indicated a measurement error of about ±0.2 dB in the magnitude measurement.

Several WR-187 band backshorts were built and tested. The varied parameters were the (a) shape of the holes, (b) size of the holes, (c) spacing of the holes, (d) number of holes, (e) thickness, T, of the backshorts and, hence, width of the gap between the backshort and the waveguide wall, and (f) thickness of the mylar insulator. Each of these parameters can affect the electrical length of the high- and low-impedance sections which determines the performance of the backshort.

Millimeter wave tests were also made in WR-10 waveguide at 75-110 GHz. The test apparatus is shown in Fig. 3. A Micro-Now backward wave oscillator (BWO) and Singer sweeper were used to provide a 75-110 GHz swept signal. A direct detector, 10 dB directional coupler, and Wiltron 560A Scalar Network Analyzer were used to detect the reflected power. The system was calibrated by placing a copper plate at the position of the reference plane at the waveguide flange.

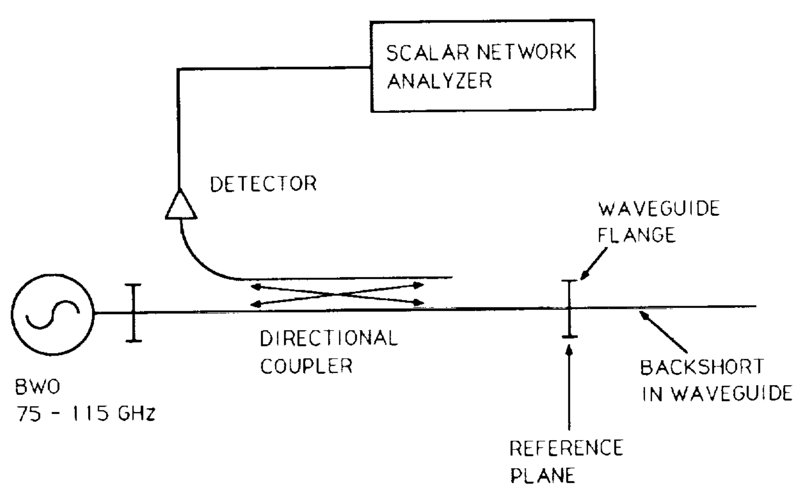


Figure 3. Millimeter wave test apparatus.

RESULTS

WR-187 Band Measurements

Figure 4 shows the reflected signal for a solid bar without holes. This backshort has dimensions W = 47.5 mm and T = 19.7 mm. This leaves a gap of 1.2 mm on either side of the bar and waveguide wall. This is a large gap, but it corresponds to typical machining tolerances to be expected for much smaller waveguides at 200-300 GHz. The mylar is 0.89 mm thick (the mylar thicknesses used for the various backshort tests were obtained by stacking two to five layers of 0.127 mm and 0.254 mm thick sheets). As seen in Fig. 4, the solid bar without holes does not make a good backshort. There are several frequency bands where the reflection is much less than -1.0 dB (0.79 reflected power). A mode is readily generated in the mylar-filled gap between the bar and the waveguide wall, and the power escapes out the end of the waveguide. At a few frequencies, the reflection coefficient approaches -0.25 dB but only over a very narrow bandwidth.

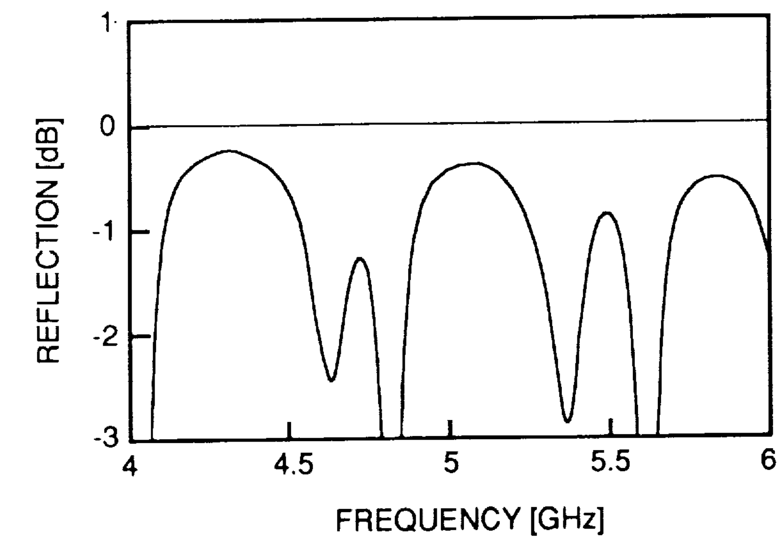


Figure 4. Reflected Power from a solid bar without holes. This design does not make a good backshort. Several large dropouts occur across the frequency band.

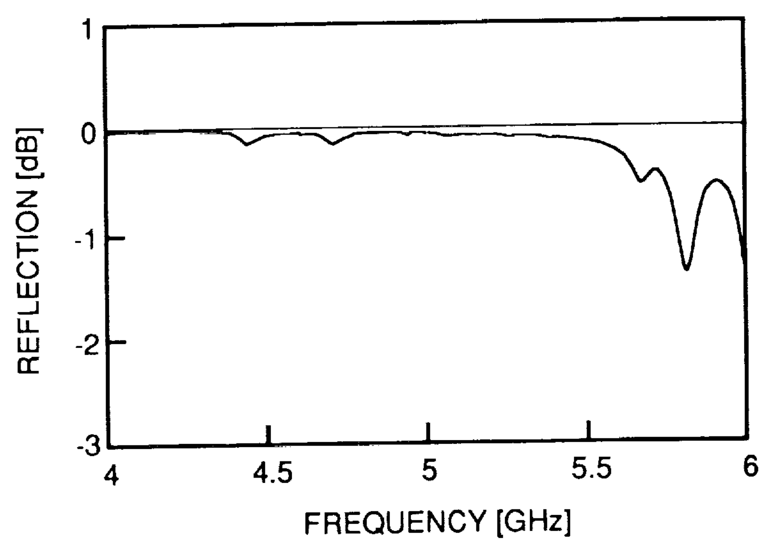


Figure 5. Reflected power from a backshort with three rectangular holes. The mylar is 0.89 mm thick. Excellent performance is obtained over a broad bandwidth.

Figure 5 shows the reflected signal for a bar with rectangular holes. The reflection coefficient is greater than -0.2 dB (0.95 reflected power) over a 33% bandwidth centered around 4.8 GHz. This is, of course, a dramatic improvement over the solid bar without holes. The backshort dimensions are W = 47.5 mm × T = 19.7 mm, and the mylar thickness is 0.89 mm. There are three holes, each with dimensions L_1 = 19.3 mm, L_2 = 28.4 mm and spacing S = 8.7 mm. Taking the center frequency of the stop band to be 4.8 GHz implies that the high-impedance section lengths are L_1 = 0.24 λg where λg = 79.1 mm, and the low-impedance sections are S = 0.17 λg where λg = 50.3 mm. The presence of the mylar modifies the waveguide modes. The guide wavelengths, λg, for the high- and low-impedance sections were calculated using a transverse mode technique which is described in references [1] and [2]. We are currently working on a full theoretical description which will allow one to calculate and design the center frequency, bandwidth, and reflection coefficient as a function of hole size, shape, spacing, and dielectric thickness [3].

A significant decrease in the reflection coefficient (a "dropout") is seen near 5.8 GHz in Fig. 5. The position of this dropout was dependent on mylar thickness. Increasing the mylar thickness, which decreases the guide wavelength, moved the dropout to lower frequency. Decreasing the thickness moved it to higher frequency. Figure 6 shows the result for a mylar thickness of 0.64 mm. The response is much flatter except for a slight decrease in reflection near 4.8 GHz. This response is comparable to that obtained for the conventional type of backshort shown in Fig. 1.

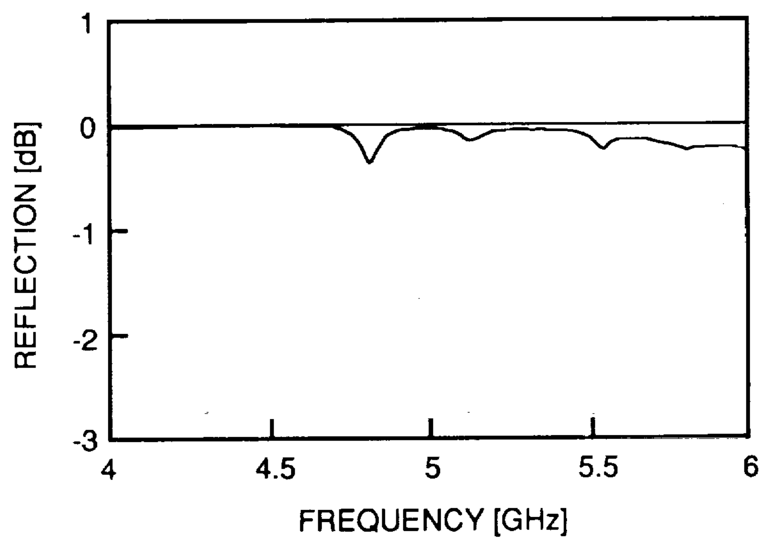


Figure 6. Reflected power from a backshort with three rectangular holes. Reducing the mylar thickness to 0.64 mm moved the dropout seen in Fig. 5 out of band. The performance is comparable to a conventional noncontacting backshort.

Figure 7 shows the reflection coefficient for a backshort with three circular holes of diameter D = 19.3 mm and spacing S = 8.7 mm. The bar dimensions are W = 47.5 mm × T = 19.7 mm, and the mylar thickness is 0.89 mm. The reflection is greater than 0.2 dB over a 32% bandwidth with a center frequency near 4.75 GHz. This gives the high-impedance section lengths D = 0.24 λg where λg = 80.6 mm, and the low-impedance section lengths S = 0.17 λg where λg = 51 mm [1, 2]. These results are similar to those obtained with the rectangular holes. This is encouraging since round holes are easier to fabricate. The dropout near 5.7 GHz is somewhat larger than that seen in Fig. 5. This probably results, in part, from power which leaks around the edges of the holes. Deceasing the mylar thickness to 0.64 mm moves the dropout to high frequency. The result is shown in Fig. 8. A new dropout, however, is beginning to appear at the low frequency end.

Many other variations of the backshort parameters, other than those discussed here, were tested. These variations affected the magnitude, phase, and bandwidth of the rf reflection. A more extensive discussion of these systematic variations will be given at a later date. The results discussed here are typical of the best performance to date.

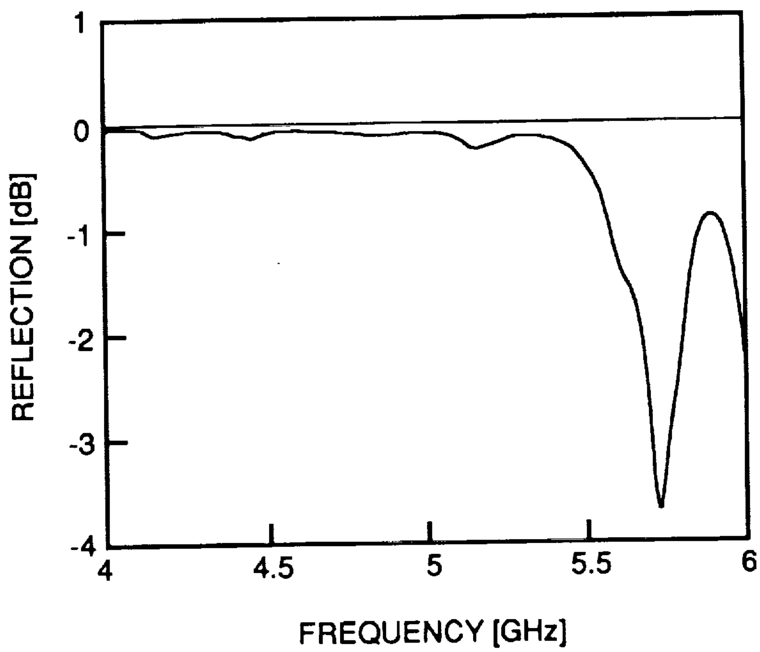


Figure 7. Reflected power for a backshort with three round holes. The mylar thickness is 0.89 mm. The performance is comparable to the backshort with rectangular holes.

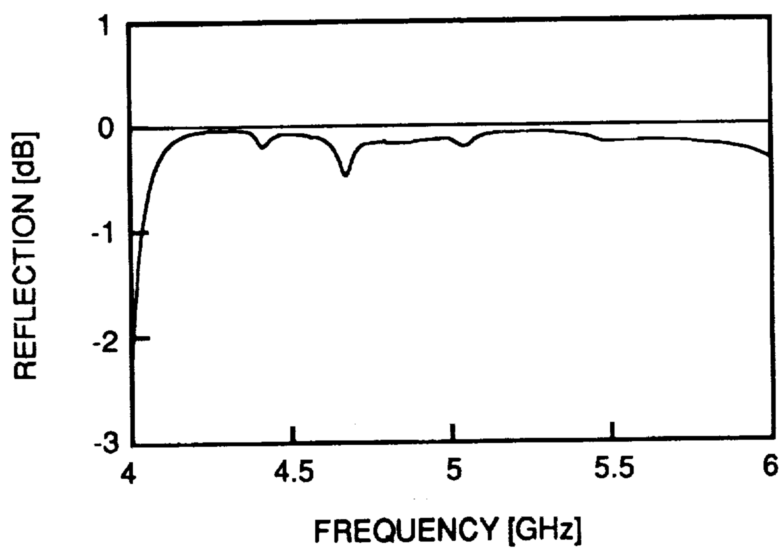


Figure 8. Same backshort as in Fig. 7, except mylar thickness is reduced to 0.64 mm. The dropout at the high-frequency end has moved out of band.

WR-10 Band Measurements

A crucial test of this new design is to measure its performance at millimeter wave frequencies. The WR-187 band backshorts were scaled for use at WR-10 band. The scale factor is 0.0535. Thus, the backshort dimensions are W = 2.54 mm × T = 1.05 mm. The WR-10 waveguide dimensions are 2.54 mm × 1.27 mm (0.10 in × 0.05 in). The frequency range, 4 GHz - 6 GHz, scales up to 75 GHz - 112 GHz.

Figure 9 shows the reflection coefficient versus frequency for a backshort with three rectangular holes. The hole dimensions and spacing were scaled from the low frequency case. The mylar is 0.051 mm thick which corresponds to 0.95 mm at WR-187 band. Thus, the results in Fig. 9 should correspond approximately to those shown in Fig. 5. As seen in Fig. 9, the performance is excellent and corresponds well with the low-frequency case. The decrease in reflection near 110 GHz corresponds almost exactly to the dropout seen near 5.8 GHz. The reflection coefficient is -0.05 dB to -0.3 dB over about a 30% bandwidth. This is suitable for practical applications. The missing sections of the curves in Fig. 9 correspond to frequencies at which the BWO was unstable and the data could not be adequately normalized.

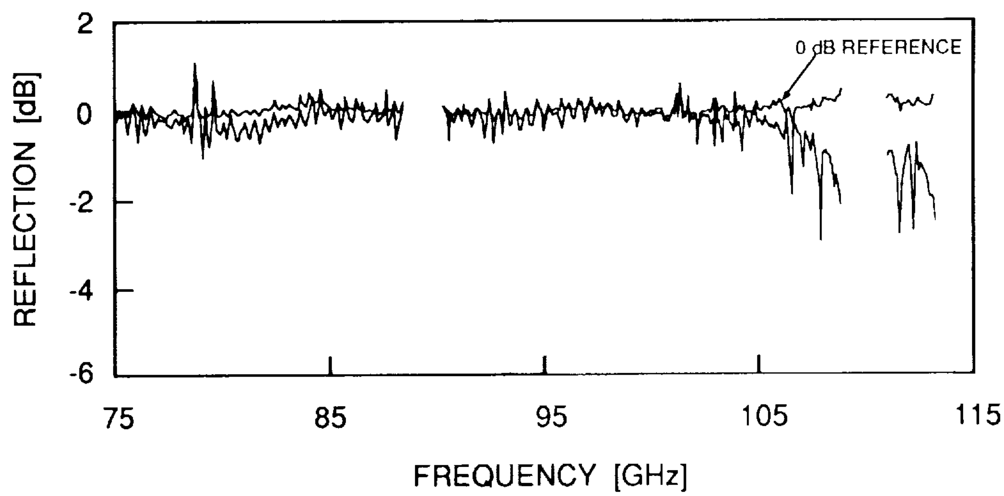


Figure 9. **Reflected power for a backshort with three rectangular holes in WR-10 waveguide. The 0 dB reference is provided by a metal plate inserted between the waveguide flanges. Excellent performance is obtained over a broad bandwidth. This result is comparable to the low-frequency case (see Fig. 5).**

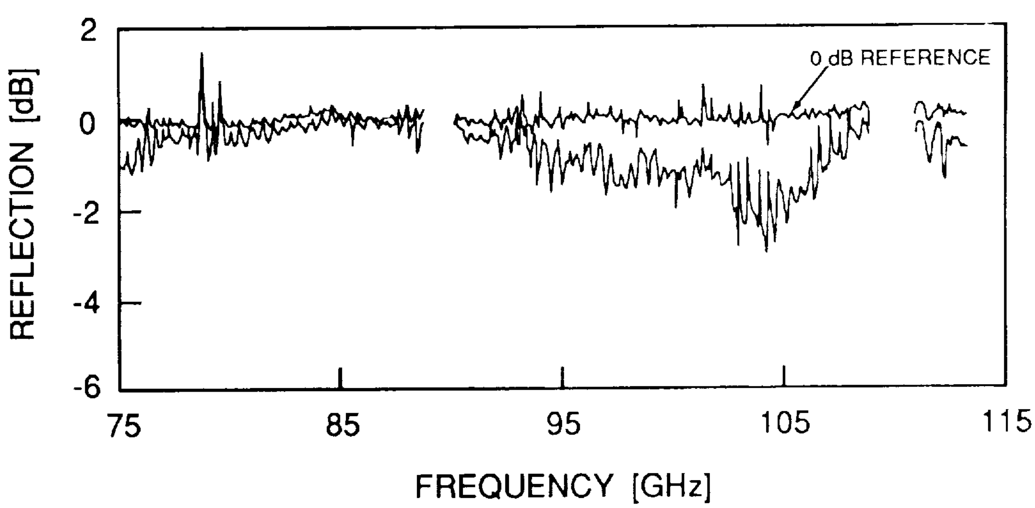


Figure 10. **Reflected power for a backshort with three round holes in WR-10 waveguide. The 0 dB reference is provided by a metal plate inserted between the waveguide flanges. The performance is comparable to the low-frequency case (see Fig. 7), but the bandwidth is narrower.**

Figure 10 shows the performance for a backshort with three round holes. These results correspond to the low-frequency case shown in Fig. 7. The reflection coefficient is -0.3 dB or better over the frequency range from about 76 GHz to 90 GHz. Again, this is well suited for many applications. However, this is only about half the bandwidth observed in the low-frequency case. The dropout near 105 GHz, nonetheless, corresponds well to that seen in Fig. 7. The generally low reflection, ≈ -1 dB, between 90 GHz and 105 GHz, however, is not seen in the low-frequency case. It may simply result from the mylar thickness not being exactly correct. Further tests with the low-frequency backshort are needed to check this as well as sensitivity to other dimensional tolerances.

CONCLUSIONS

A new noncontacting waveguide backshort design has been developed which provides performance as good as the more developed conventional approaches. It employs a metallic bar with rectangular or circular holes to provide a periodic variation of the waveguide impedance on the correct length scale to result in a large reflection of rf power. This design is mechanically rugged and can be easily fabricated using a variety of methods for frequencies from 1 GHz to 1000 GHz. It should allow tunable waveguide systems to be extended well above 300 GHz.

ACKNOWLEDGEMENTS

We wish to thank M. A. Frerking and P. Siegel for valuable discussions. This work was supported in part by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration and the Innovative Science and Technology Office of the Strategic Defense Initiative Organization.

REFERENCE

[1] M. K. Brewer and A. V. Raisanen, *IEEE Trans. Microwave Theory Tech.* **MTT-30**, 708 (1982).

[2] R. F. Harrington, Time-Harmonic Electromagnetic Fields, McGraw-Hill, New York, pp 158-161 (1961).

[3] T. Weller, P. B. Katehi, and W. R. McGrath, to be published.

MATERIALS SCIENCE

(Session A6/Room A2)

Tuesday December 3, 1991

- Novel Applications of TAZ-8A
- Test Methods for Determining the Suitability of Metal Alloys for Use in Oxygen-Enriched Environments
- A Major Advance in Powder Metallurgy
- Permanent Magnet Design Methodology

N92-22444

NOVEL APPLICATIONS FOR TAZ-8A

Stephen M. Riddlebaugh
William J. Waters
NASA Lewis Research Center
Cleveland, OH 44135

ABSTRACT

Since the early 1960's the NASA Lewis Research Center has been actively engaged in alloy development for jet engine applications. A series of alloys were developed through "in-house" research. Several significant alloys resulted from this program. One of the most promising alloys is a nickel base material referred to as TAZ-8A. It was selected by NASA as the nose cone material for the Mach 8 version of the X-15 rocket ship.

Recent needs in the non-aerospace industrial sector have revitalized interest in high performance alloys. TAZ-8A has a combination of properties that makes it unique: (a) a high temperature strength, (b) oxidation resistance, (c) abrasion resistance, and (d) exceptional thermal shock resistance. The major drawback for utilization of this alloy is the relatively high cost compared to the more common iron base alloys. Reduced material consumption and hence lowered costs are possible by using coatings of TAZ-8A on a low cost substrate.

Coatings have been applied using plasma spray techniques developed by NASA as well as modified plasma vapor deposition (PVD) techniques.

Unique properties result from each of these two different coating processes. The PVD process results in a thin coherent coating that possesses high reflectivity, extreme hardness, and abrasion resistance. These properties are currently being quantified and offer the potential for a wide variety of commercial applications.

INTRODUCTION

To a large extent the history of the development of gas turbine engines has been a story of the development of high performance materials. The same is true for all "heat cycle" engines. Virtually all of these engines convert thermal energy to mechanical work by heating a working fluid in a controlled volume such that as it expands it exerts pressure against a surface, translating it into a force acting on the machine. The primary source of the heat used in these engines is the heat energy released in the combustion of a fuel. Much of this heat energy is wasted through cooling systems or by exhausting the hot working fluid prematurely, because the materials and lubricants in the engine connot withstand the elevated operating temperatures that would be necessary to fully utilize all of the heat energy generated. As a consequence, scientists and engineers have been working for three hundred years to develop high temperature materials.

In the 1960's scientists, engineers and technicians at the NASA Lewis Research Center developed a family of nickel based "superalloys" employing tantalum as the principal alloying ingredient. The most promising of these was an alloy containing eight weight percent tantalum, called TAZ-8. Further development of the TAZ-8 alloy to improve oxidation resistance produced TAZ-8A, in which columbium was substituted for vanadium in the original alloy (1). The Lewis-developed TAZ alloys became the prototypes of a number of similar high temperature superalloys developed by the aerospace industry.

While the initial "target applications" of TAZ-8A were hot section components of gas turbine engines (such as turbine vanes or "buckets"), other aerospace applications soon appeared. One example is the nose cone of the X-15 rocket plane. In this case the source of the high temperature was the aerodynamic frictional heat buildup at Mach 8 flight.

There are some drawbacks to the utilization of this superalloy. One is its toughness. It is extremely resistant to the normal cutting and grinding processes normally used to machine metal parts to their final shapes. The material can be worked with forming processes based on plastic deformation of the material, such as rolling, but only at very small deformation rates and under very precisely controlled conditions. Considerable work had to be done to develop the necessary manufacturing technologies to use this material. The two principal techniques used are precision casting to near net shape and the consolidation of prealloyed powder.

The other drawback is cost. Some of the alloying elements are "strategic materials", meaning they are rare, difficult to obtain in quantity, and thus expensive. Besides raising the cost, this limits the size of the components and their number. The solution to this problem is to use TAZ-8A as a thin, protective coating on a low cost substrate. This will also bypass many of the difficulties encountered in fabricating complex parts.

It is doubtful that TAZ-8A would be a practical material for use in mass produced consumer products, at the present. Rather it is felt that its major potential at the present time is to improve the performance of key components in industrial processes involving extremely high temperatures, continuous thermal cycling, and abrasive environments. In recent years many American industries utilizing such processes have come under intense competition from abroad. This competition is often in the form of higher productivity rates, leading to lower prices of the final products. One way to improve productivity is to reduce down time by identifying the critical components, the "weak links", and increasing their service lives.

PROPERTIES OF TAZ-8A SUPERALLOY

The nominal composition of the TAZ-8A alloy in weight percent is shown in Table 1. The relative amounts of the constituent elements were determined in an iterative process to optimize them. For example, columbium content was varied from 0.5 to 20 percent, but it was found that stress-rupture life reached its maximum at 2.5 percent and degraded rapidly as the columbium content was increased beyond that level(1).

It should be noted that the manufacturing processes and test conditions used to evaluate TAZ-8A were chosen as characteristic of its anticipated aerospace applications, primarily as a material from which to fabricate hot section components of aircraft gas turbine engines. The initial work was done with investment-cast specimens melted either in an inert gas (argon) furnace or a vacuum furnace. Specimens cast included stress-rupture and tensile test bars, Charpy impact test bars, and blanks for test rolling sheet strips to test workability.

Figures 1 through 7 present high-temperature properties of cast TAZ-8A compared to representative examples of other high-temperature materials used for similar aerospace applications. Representative example data are used rather than data for specific alloys because the aerospace industry has developed a number of different alloys in each generic category, and each alloy has slightly different properties. For the purpose of this paper, "typical" properties for each category are more instructive.

Figures 1 and 2 present stress-rupture data taken at 1800F. The temperature was selected as representative of gas turbine hot-section temperatures at the time the tests were performed. Stress-rupture data addresses the problem of how well the material will hold up under load in a high temperature environment. This is a more realistic approximation of what a machine element experiences than a tensile test.

TAZ-8A is a cast material, and Figure 1 shows what a major impact that controlled solidification of the molten alloy can have on the material's strength. Stress-rupture performance is improved by nearly 40 percent by directionally solidifying the melt. Considerable care is required in the casting process with a highly alloyed material such as TAZ-8A to prevent segregation (the separating-out of the constituents) during solidification, resulting in a part with varying properties.

Figures 3 and 4 present high temperature tensile strength data. The proportional limit is the stress level at which ductile deformation of the test sample initiates. The ultimate tensile strength is the stress level at which the piece fails. Notice that directional solidification reduces the ultimate tensile strength of TAZ-8A, and also lowers the yield point, but at the same time significantly increases its ductility. This would produce a machine part that would be less prone to catastrophic failure.

The number of thermal cycles a standard test bar endures before cracking initiates on its edge is a measure of the material's ability to withstand thermal shocking and repeated hot-cold temperature cycles. Thermal cycling data for TAZ-8A compared to other high temperature alloys are presented in Figures 5 and 6. This material exhibited outstanding thermal cycling resistance in the testing that produced these data. The temperature ranges were selected as representative of aerospace conditions.

Oxidation resistance is another important property for materials used in high temperature applications. Oxidation behavior of TAZ-8A is compared to other representative high-temperature alloys in Fig. 7. Fig. 7a presents weight-gain data after 8 hours exposure at several temperatures. The increase in weight is due to the oxygen takeup by the outer layers as they oxidize. This oxidized scale was then removed; the net weight loss is presented in Fig. 7b.

The TAZ-8A specimens used for the oxidation data in Fig. 7 were vacuum-melted. With samples cast in an inert gas (argon) atmosphere, considerably poorer oxidation resistance was measured, although overall oxidation resistance was judged comparable to other high temperature materials. Table 2 compares oxidation behavior of vacuum-melted and argon-melted TAZ-8A. The difference in oxidation rate appears to be due to a coarser microstructure in the argon-melted material. Photomicrographs of the metal-oxide interface, depletion zone, and unaffected matrix of oxidised stamples of vacuum-melted and argon-melted TAZ-8A showed a thinner depletion zone with a more clearly defined interface between the depletion zone and the unaffected matrix in the vacuum-melted specimens. As in the case of the tensile data, this shows that the processing and fabrication procedures can have a significant impact on the performance of the finished part.

A limited workability potential of the cast TAZ-8A material was demonstrated by hot-rolling thin (0.110 in.) cast strips into sheet strips approximately 0.020 in. thick(1). However, the process used was characterized as being "somewhat specialized" and concluded that TAZ-8A was not "demonstrated to be a wrought alloy." In addition, TAZ-8A is a very hard, tough alloy that resists cutting and grinding operations unless diamond-edge cutting tools are used. The most practical method of fabricating parts of cast TAZ-8A is to precision cast the part to near-net shape and finish-grind it if necessary.
Because the material is a highly alloyed one, the casting process must be carefully controlled to prevent segregation of the constituents during solidification.

As part of the effort to reduce the workability and casting difficulties encountered with TAZ-8A, an investigation was made into extruding bars of the material from prealloyed powders(2). The powder was made by atomizing the molten alloy in an inert gas spray. The fine droplets of molten alloy were subjected to very rapid solidification, preventing segregation. The powders were sealed into evacuated mild steel cans and extruded directly into bars. The resulting bars had a very fine, homogeneous microstructure.

The extruded bars exhibited a curious reversal of properties compared to the cast TAZ-8A material. At temperatures below 1500F (830C) the extruded bars had considerably greater tensile strength than the cast bars, but at higher temperatures the cast bars were stronger. Heat treating the extruded bars to cause grain coarsening resulted in bars with tensile and stress-rupture properties that approached the as-cast bars but did not match them. The strength and stress-rupture properties of the extruded bars, compared to the cast bars, are presented in Figures 8, 9, and 10.

Notice that the stress-rupture data in Fig. 10 is for heat-treated extruded bars. Stress-rupture data were unobtainable for the as-extruded bars because under the controlled conditions of the stress-rupture test they exhibited superplastic behavior. After four hours at 1900F (1038C) and 1000 psi (6.89 MN/square meter) the as-extruded TAZ-8A specimen had an elongation of over 600 percent without rupturing. At that point the test apparatus was at its maximum travel.

This superplastic behavior suggests one approach to forming parts from TAZ-8A. In a subsequent test involving hot pressing the extruded powder material under low strain rates a cylindrical specimen 9/16 in. (1.4 cm) diameter and 5/8 in. (1.6 cm) high was reduced in height by 75 percent without cracking(2). However, when attempts were made to form the material at high strain rates, such as are encountered in conventional rolling and stamping operations, the material cracked. These results indicate that bars extruded from TAZ-8A powder can be "wrought" providing slow strain (deformation) rates are used and the formed product then heat-treated to "set" it. The heat-treated extruded bars did not exhibit superplastic behavior; they behaved similarly to the cast specimens.

The high cost of TAZ-8A alloy compared to the common iron base alloys, combined with the rather specialized fabrication procedures required to work this material, limit its use to relatively small parts if they are made of the solid alloy. This has led to investigations of using the material as a high-temperature protective coating applied to less expensive substrates. Used in this manner its properties of exceptional thermal shock resistance, oxidation resistance and abrasion resistance can be transmitted to large complex machine components while performing the difficult machining work on the more easily worked and less expensive substrate.

At the present, two techniques for applying TAZ-8A coatings to various materials have been investigated: plasma spray coating and plasma vapor deposition (PVD). The resulting coatings have not been fully characterized, but some preliminary results can be reported.

Plasma spray techniques developed by NASA can deposit layers of TAZ-8A up to about 20 mils thick. When thicker coatings were attempted, the plasma spray became unstable and the coating would separate from the substrate. The resulting surface is rough but can be ground smooth. The plasma sprayed layers are porous and include metal oxides dispersed throughout the coating layer (a consequence of spraying molten alloy in air). It has not been quantitatively determined what effects these encluded oxides have on the properties of the coating. It may be that for certain applications they may have a beneficial effect; for example, they might improve the abrasion resistance of the coating. In test applications so far, these oxides do not appear to adversely effect the performance of the coating.

The porosity of the plasma spray coating means that this material will not protect the substrate material from corrosion. In one preliminary test of a steel component with a plasma spray coating of TAZ-8A, subjected to a very high temperature environment with water spray cooling, the water penetrated to the base metal and corroded it, causing the TAZ-8A layer to spall off. At the time of coating separation, the test part had lasted in service ten times as long as the standard uncoated part.

Technicians at the NASA Lewis Research Center have plasma vapor deposited TAZ-8A coatings on a variety of materials, including stainless steel, fiberglass, glass, and ceramic materials such as Space Shuttle thermal protection tiles. PVD coatings up to 2 mils thick have been applied; since the PVD process is time-dependent, thicker coatings would not be economical by this process.

The PVD coatings produced so far have been observed to be clean, very dense, with a very fine microstructure. The grains are very thin, closely packed, and aligned perpendicular to the surface being coated. They appear to be very high strength in the direction perpendicular to the surface, but there is some preliminary evidence that the coating layer may readily "cleave" along grain boundaries. This may be prevented by heat treatment to realign the grain structure and coarsen it. Preliminary results also indicate these PVD coatings have excellent tribological properties. When applied to smooth surfaces these PVD coatings appear to be highly reflective, suggesting their possible use as a high temperature, corrosion-resistant, abrasion-resistant mirror surface. Surface imperfections appear to be transmitted through the coating with no "blending out," suggesting use as a coating for die casting dies.

COMMERCIAL APPLICATIONS OF TAZ-8A

The Technology Utilization Office at the NASA Lewis Research Center is engaged in several Technology Applications Projects in which TAZ-8A is being investigated as a solution to problems inherent in some high temperature industrial process. These specific projects are still in preliminary stages, but they offer insights into the technical needs of such industries, and thus into potential applications of materials such as TAZ-8A. In addition, properties of this material in its various forms (cast, extruded powder, plasma spray or PVD coatings) that have been investigated also suggest potential applications.

The aerospace industry routinely casts superalloy materials into complicated, precision shapes such as turbine blades for jet engines. Precision casting techniques such as investment casting appear to be the best way to fabricate solid machine elements. Cast TAZ-8A alloys are extremely tough and resistant to normal cutting-type machining operations such as milling. Grinding and abrasive cutting using diamond tools are required.

This hardness, while a liability during the fabrication of a part, may become an asset, as the component will be highly abrasion resistant. A number of industrial process environments are both high-temperature and very abrasive. An example is glass making. The "chill blocks" used to cut glass sheets and tubes in several processes might be likely candidates for making from TAZ-8A.

The cost of the material and the casting processes (inert gas or vacuum casting) at present would limit the size of such parts to fairly small sizes. Some larger pieces could be fabricated from extruded powder bars by superplastic deformation, although the slow strain rates required for this will result in long fabrication times.

Thermal fatigue failures often propagate from surface cracking or crazing at the point of exposure to high temperature; this could be prevented or at least delayed by applying a high temperature protective coating. Using TAZ-8A as a coating applied to components fabricated from lower cost, more easily machined materials reduces the amount of TAZ-8A needed, eliminates many of the fabrication problems, and permits using this material on large parts. Advanced ceramics and ceramic composites have been proposed for similar applications. Superalloys such as TAZ-8A still appear to be more physically tough and resiliant, and able to withstand considerably more and more extreme thermal cycling and shocking than ceramic materials.

Several ongoing applications projects being conducted by the Lewis Technology Utilization Office are using this approach to attempt to solve thermal fatique problems in the machinery used for the continuous casting of steel. Also, the abrasion resistance and thermal shock resistance of TAZ-8A has led to an investigation into seeing if it could solve problems of die wear and shot sleeve erosion in the die casting industry.

To date, the two coating techniques investigated are plasma spray coating in air and plasma vapor deposition in a reduced atmosphere. The plasma spray coating process can produce a thicker coating, but the surface is rough and has to be ground if a smooth surface is required. This limits surface shapes to those that can be ground. The coating is porous; if the operating environment can attack and corrode the substrate material, some sort of impervious seal coat would be necessary. The PVD technique produces a very dense, impervious coat of TAZ-8A; this makes it a candidate seal coat. PVD also is the most promising technique for coating die casting dies, as it lays down a very smooth surface that replicates all details of the substrate surface.

A wide variety of materials has been successfully coated with TAZ-8A using the PVD process. Among them are stainless steel, glass, and ceramics such as the Space Shuttle tiles. It is felt that high temperature polymers such as PMR-15 are also likely candidates for coating with TAZ-8A using PVD. On smooth substrates, very smooth, highly reflective mirrorlike coatings of PVD TAZ-8A have been achieved. While the reflectivity of these TAZ-8A coatings haven't been quantitatively measured, the preliminary results suggest that this material could be used to produce mirrors for high temperature, abrasive environments.

Another potential application that as yet has not been explored is high temperature bearings. The PVD coatings appear to have very good tribological properties, and the extreme resistance of the cast alloys to machining also suggests this application.

Thicker coatings than those achieved by plasma spraying or PVD would probably have to be made by bonding a thin, extruded powder sheet of TAZ-8A to the substrate, using a bonding technique such as furnace brazing. The TAZ-8A sheet could be molded to a complex substrate shape using superplastic deformation, with subsequent heat treatment to "set" it.

The fabrication and processing techniques developed so far to work TAZ-8A do not lend themselves to mass production at the volumes associated with consumer goods. Rather, it is anticipated that realistic commercial applications of this material will be limited to solving temperature and abrasion-related problems in industrial process machinery. Each potential application will have to be carefully evaluated as to whether the increase in operational life of a critical part justifies the expense and fabrication problems.

CONCLUSIONS

The nickel base aerospace superalloy TAZ-8A has several properties that suggest its potential application to solve a number of industrial process problems. These properties include high temperature strength, oxidation resistance, thermal cycling resistance, and abrasion resistance. Applied as coatings to critical machine elements, TAZ-8A has the potential to dramatically extend the in-service lives of parts exposed to high temperature processes.

REFERENCES

1. Waters, W.J. and Freche, J.C., 1966, "Investigation of Columbium-Modified NASA TAZ-8 Superalloy," NASA Technical Note D-3597.

2. Freche, J.C., Waters, W.J. and Ashbrook, R.L., 1969, "Evaluation of Two Nickel-Base Alloys, Alloy 713C and NASA TAZ-8A, Produced by Extrusion of Prealloyed Powders," NASA Technical Note D-5248.

TABLE I. COMPOSITION OF TAZ-8A

CONSTITUENT	NOMINAL COMPOSITION, WEIGHT PERCENT
Tantalum, Ta	8
Chromium, Cr	6
Aluminum, Al	6
Molybdenum, Mo	4
Tungsten, W	4
Columbium, Cb	2.5
Zirconium, Zr	1
Carbon, C	0.125
Boron, B	0.004
Nickel, Ni	Balance

Unnotched Charpy Impact Resistance 24 ft-lb at Room Temperature

TABLE II - Comparison of Oxidation-Affected Zone Depths of Argon-Melted and Vacuum-Melted TAZ-8A

ALLOY	EXPOSURE CONDITION		EXTERNAL SCALE THICKNESS, MILS	DEPLETION ZONE THICKNESS, MILS	TOTAL AFFECTED DEPTH, MILS
	TIME, HOUR	TEMP., F			
Argon-Melted TAZ-8A	310	1900	.1	.8	.9
Vacuum-Melted TAZ-8A	310	1900	.1	.3	.4

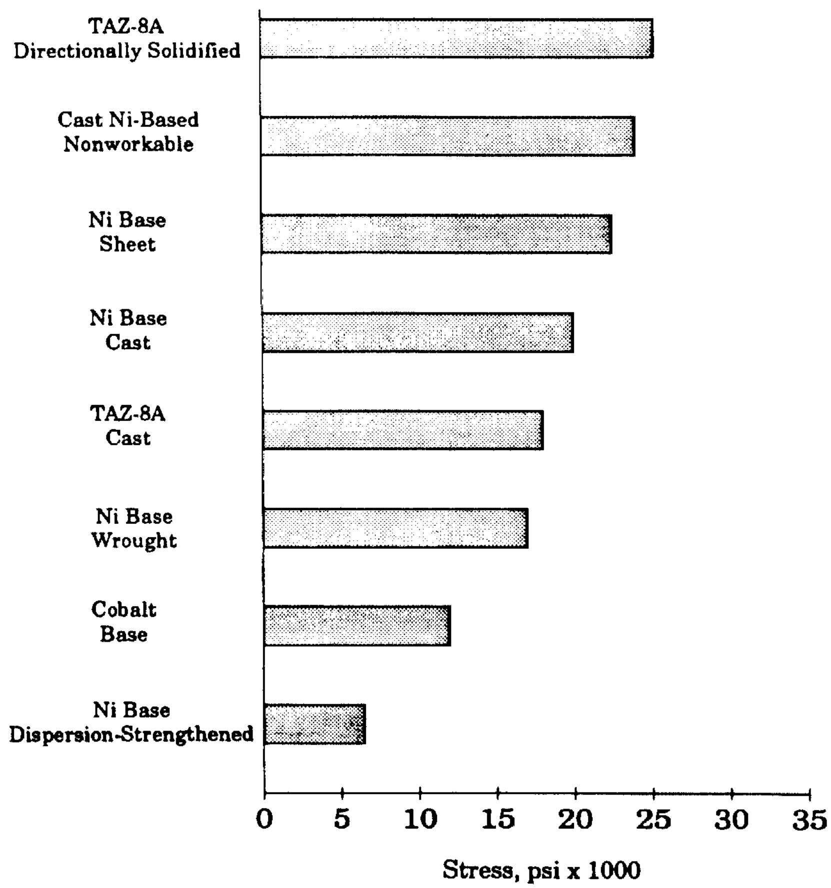


FIG. 1. - Stress to produce Rupture in 100 hours at 1800 F (1000C)

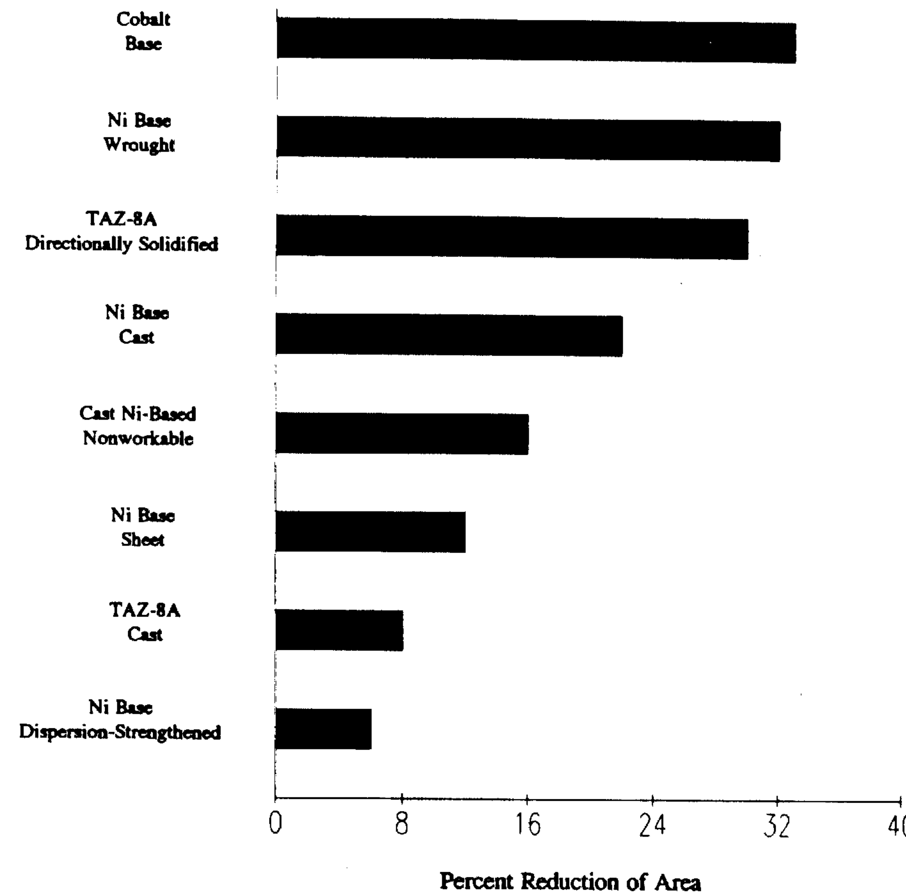


FIG. 2. - Stress-Rupture Reduction of Area at 1800 F (1000C)

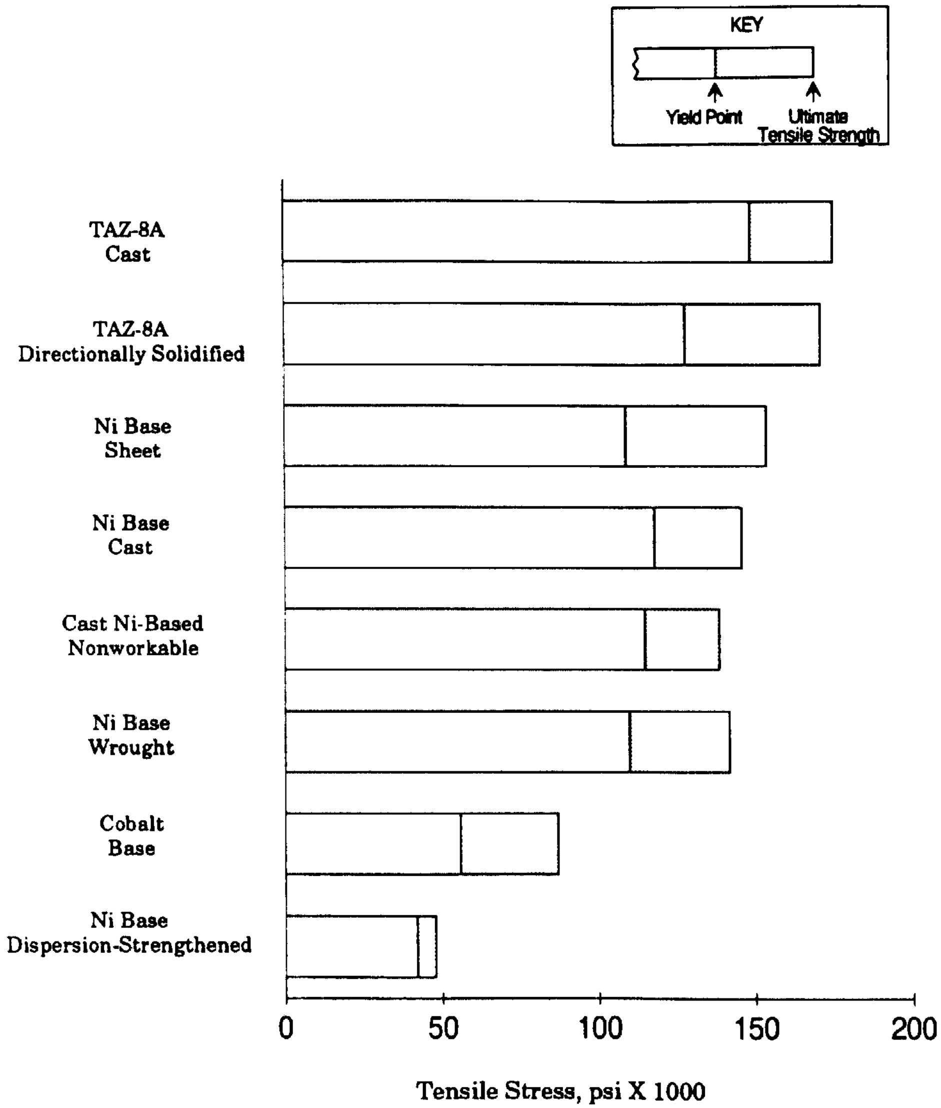


FIG 3. -Tesile Strength of Various Alloys at 1400F (775C)

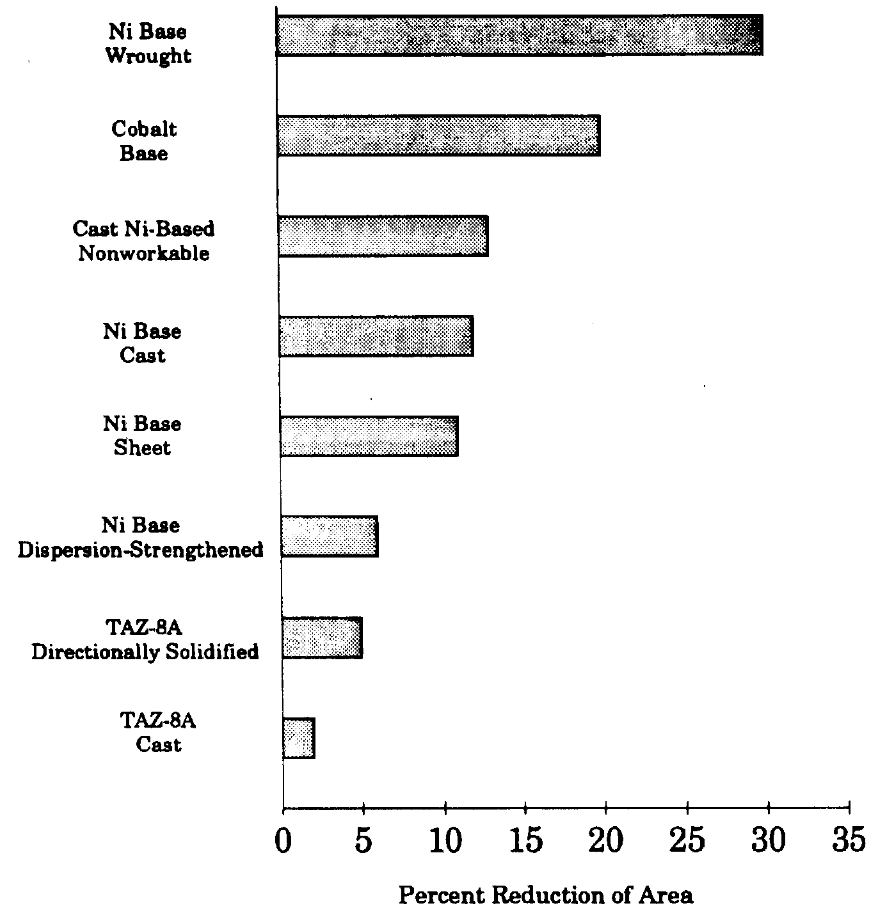


FIG. 4 - Tensile Reduction in Area at 1400F (775C)

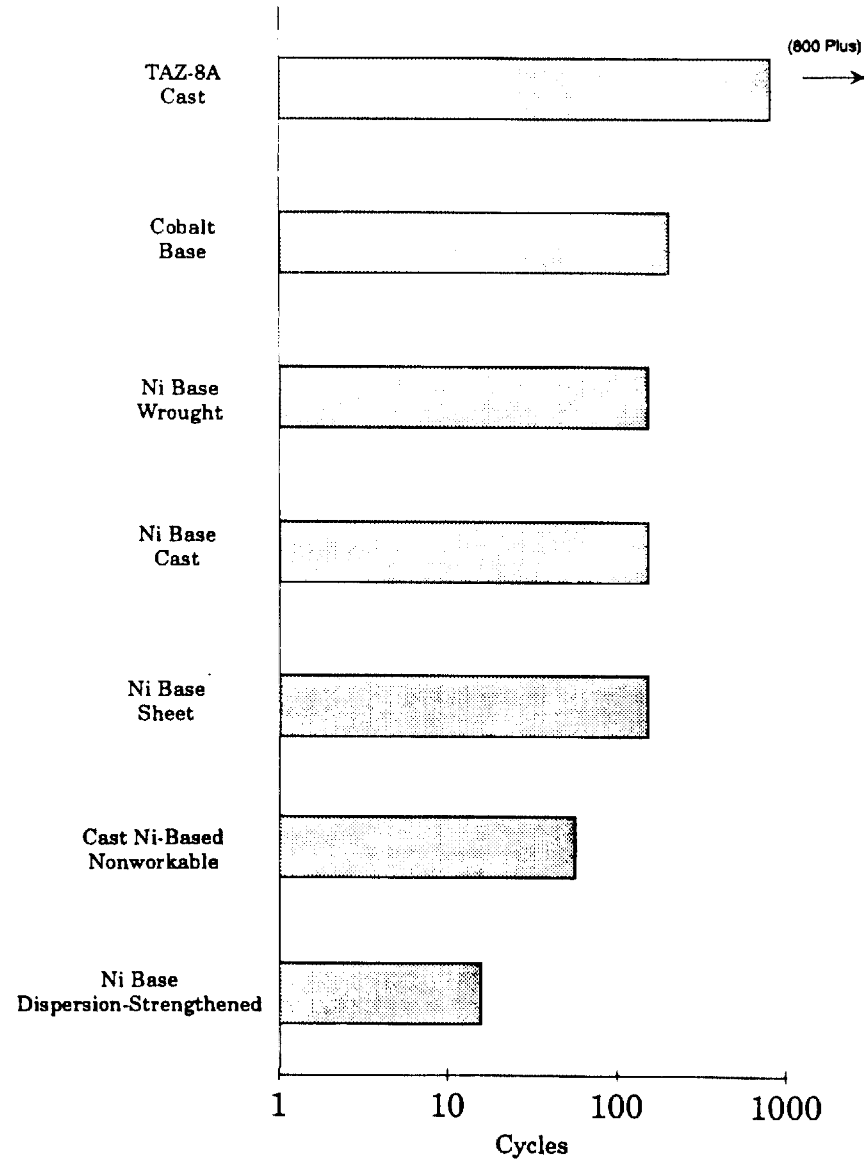


FIG. 5 - Thermal Cycles to Crack Initiation, Bed Temperatures 1915F-525F (1046C - 274C)

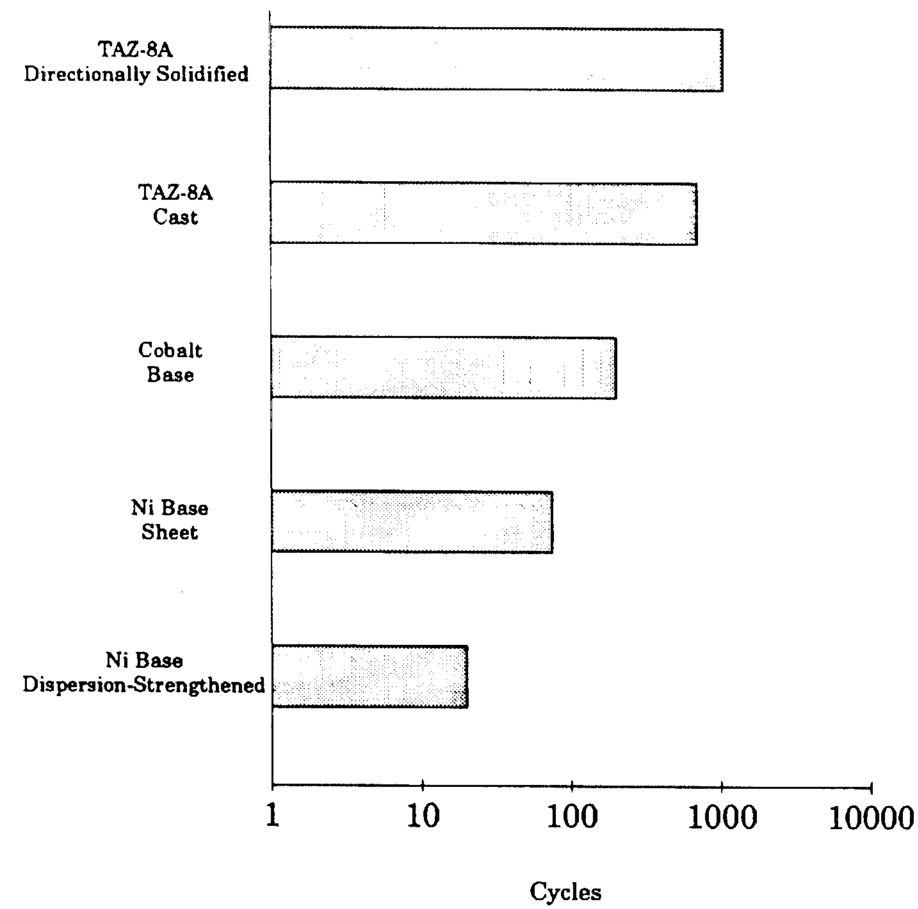
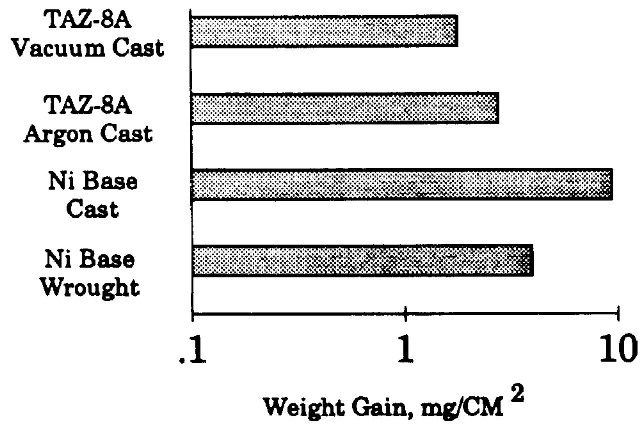
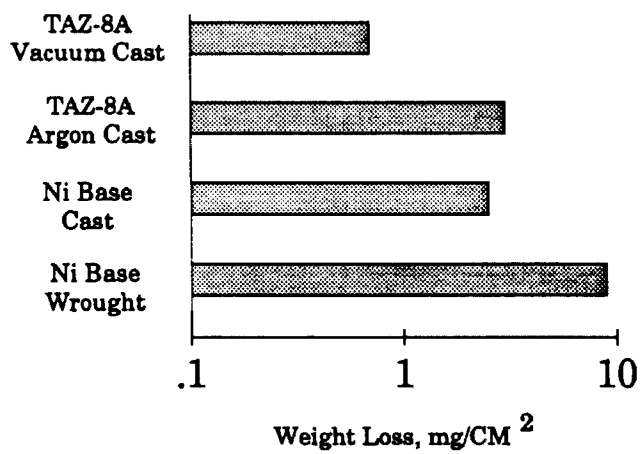


FIG. 6 - Thermal Cycles to Crack Initiation, Bed Temperatures 2065F - 675F (1129C - 357C)



(a) Weight-gain Comparison



(a) Weight-Loss Comparison

FIG. 7 - Oxidation Behavior of Several Nickel-Base Alloys at 1900F (1040C) after 310 Hours Exposure

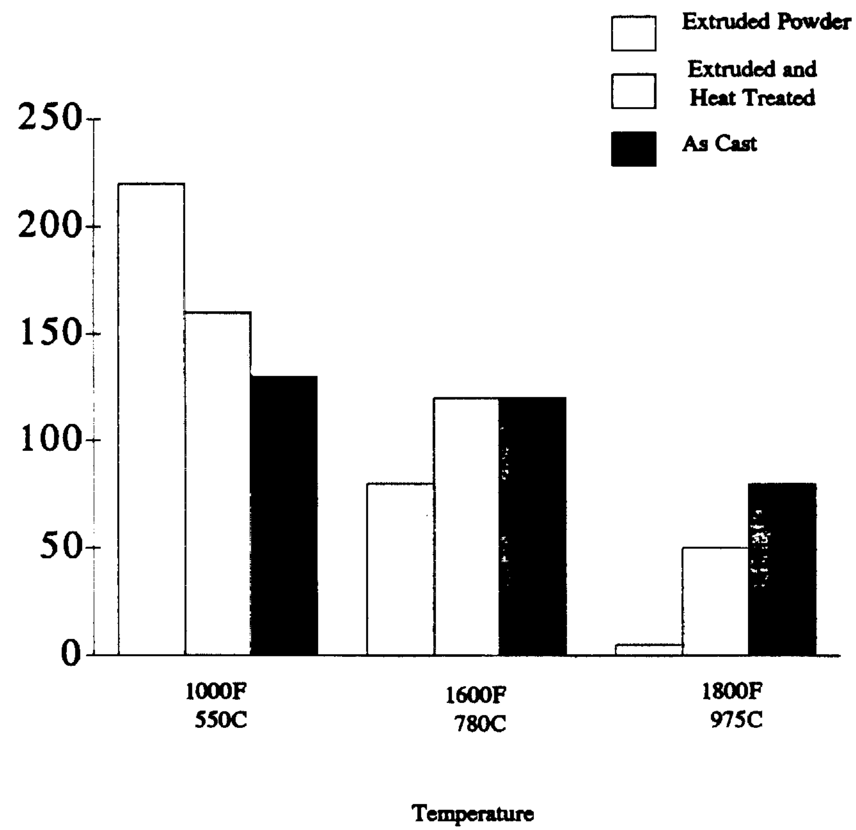


FIG 8. - Comparison of Tensile Properties of TAZ-8A Powder Products and As-Cast TAZ-8A

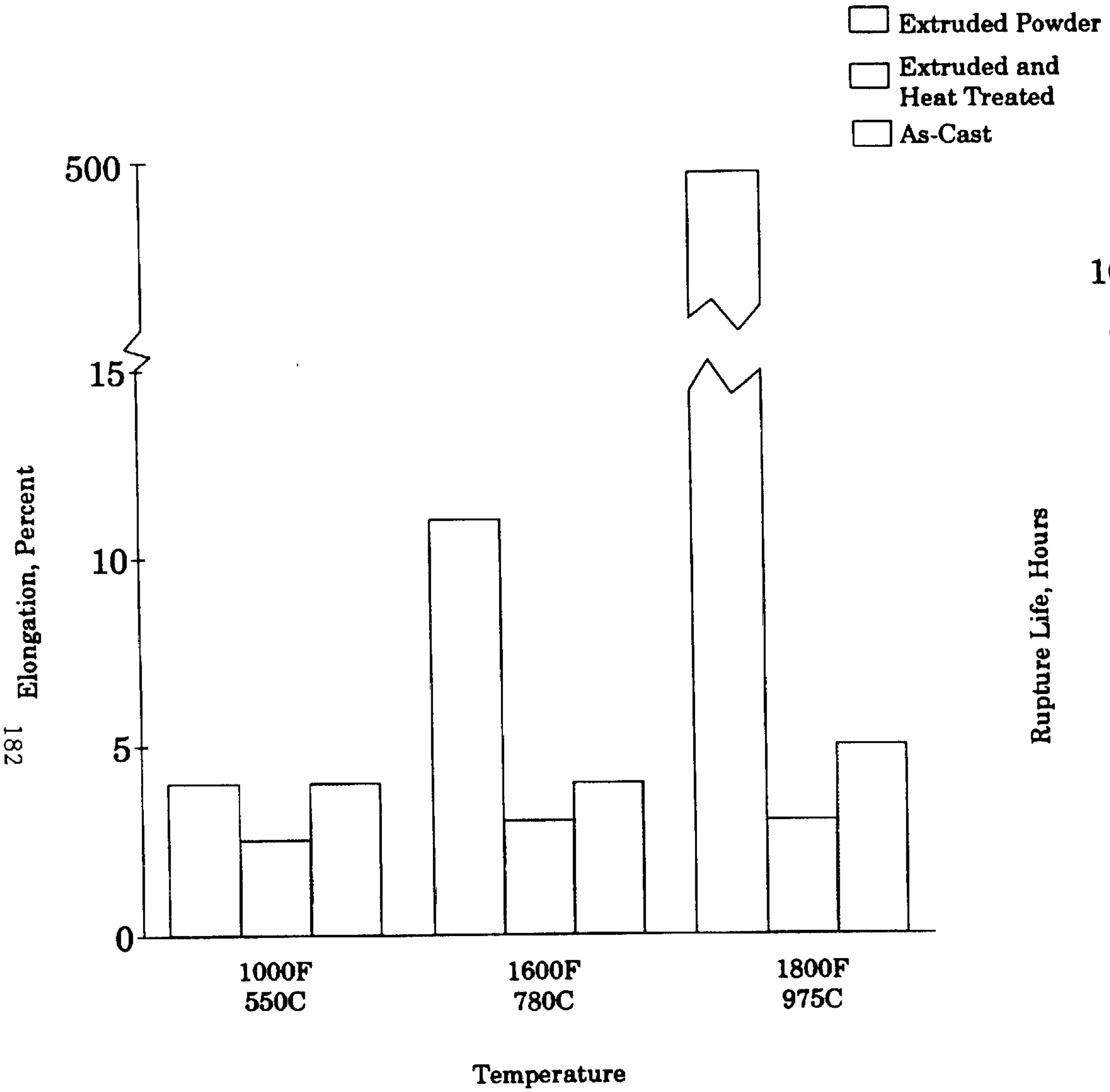


FIG. 9 - Percent Elongation at Tensile Failure of TAZ-8A Extruded Powder Products Compared to As-Cast TAZ-8A

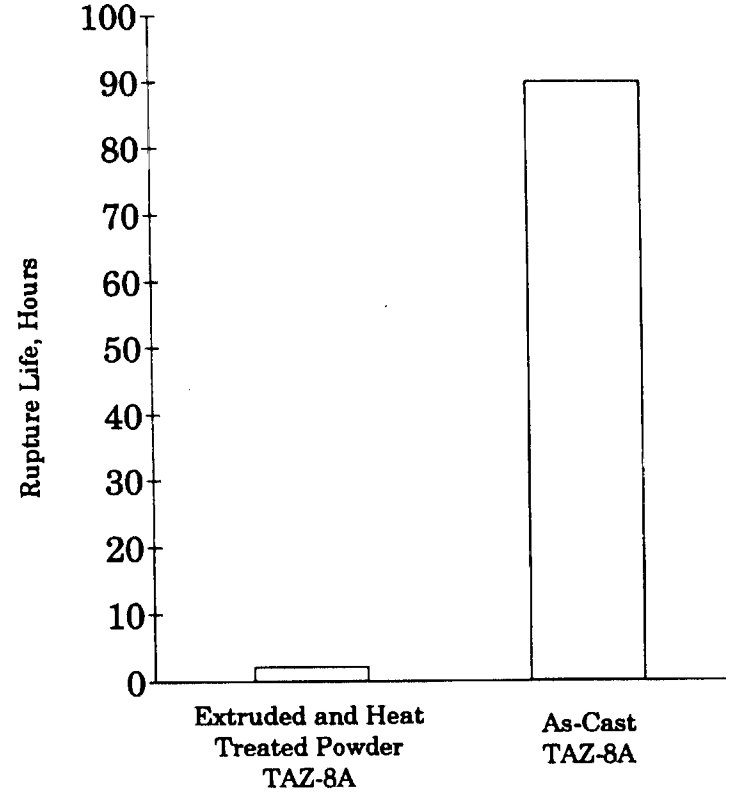


FIG. 10 - Comparison of 1900F (1038C), 15000 psi (103 MN/M^2) Rupture Lives for Extruded and Heat Treated Powder Product and As-Cast TAZ-8A

182

N92-22445

TEST METHODS FOR DETERMINING THE SUITABILITY OF METAL ALLOYS FOR USE IN OXYGEN-ENRICHED ENVIRONMENTS

Joel M. Stoltzfus
NASA White Sands Test Facility
Las Cruces NM

Mohan V. Gunaji
Lockheed Engineering and Sciences Company, Inc.
Las Cruces NM

ABSTRACT

Materials are more flammable in oxygen-enriched environments than in air. When the structural elements of a system containing oxygen ignite and burn, the results are often catastrophic, causing loss of equipment and perhaps even human lives. Therefore, selection of the proper metallic and nonmetallic materials for use in oxygen systems is extremely important. While test methods for the selection of nonmetallic materials have been available for many years, test methods for the selection of metal alloys have not been available until recently. Several test methods are presented that were developed recently at NASA's White Sands Test Facility (WSTF) to study the ignition and combustion of alloys, including the supersonic and subsonic speed particle impact tests, the frictional heating and coefficient-of-friction tests, and the promoted combustion test. These test methods are available for commercial use.

INTRODUCTION

Because nearly all metallic and non-metallic structural materials are flammable in oxygen-enriched atmospheres, costly fires have occurred in oxygen systems. For example, catastrophic oxygen-related fires have occurred in the space program, hindering mission success and costing human lives [1]. In the medical community, fires in operating rooms have occurred during surgery of the head and neck [2-5]. In industry, numerous fires have been reported in regulators and manual valves, causing severe system damage and human injury [6-8].

Furthermore, advancing technology is creating a demand for higher oxygen-use temperatures and pressures. NASA is investigating the use of liquid oxygen (LO_2) and liquid hydrogen (LH_2) as propellants for propulsion systems for a new generation of space-based orbital transfer vehicles (OTVs). The new design concept for the LH_2/LO_2 engines involves the use of gaseous oxygen (GO_2) at 533 K (500 °F) and 34.5 MPa (5000 psi) to drive the turbine in the LO_2 turbopump. Industry is considering the use of gas cylinders at up to 27.6 MPa (4000 psi). The use of oxygen at these higher temperatures and pressures causes a concern for engineers because it increases the risk of fire.

No absolute solution exists for controlling fire hazards in oxygen systems. However, the possibility of oxygen-related fires can be diminished with proper design practices and careful selection of materials. Guidance regarding proper design practices for oxygen systems and the selection of nonmetallic materials is available in the open literature [1,9-11]. Until recently, however, test methods and selection procedures for metals were not readily available.

The development of test methods for determining the suitability of metals in oxygen-enriched atmospheres began at NASA's White Sands Test Facility (WSTF) in the late 1970's. Several methods to determine the ignitability and combustion characteristics of metals have been developed. Ignition characteristics are determined using the particle impact, frictional heating, and coefficient-of-friction tests; and combustion characteristics are determined using the promoted combustion test [9,10]. This paper describes each of these test methods.

PARTICLE IMPACT TEST METHODS

Particle impact has been recognized as an ignition source in oxygen systems for several years [12-15]. Additionally, experience at WSTF has demonstrated that metal particles entrained in flowing oxygen can ignite valves, as seen in Figure 1. Two particle impact test methods, one operating at supersonic velocities and the other at subsonic velocities, have been developed at WSTF. The methods are used to determine the minimum temperature, pressure, and velocity at which particle impact ignition occurs.

Supersonic Particle Impact Test Method

The supersonic particle impact test chamber (Figure 2) has been described by Benz et al. [16] It comprises a gas inlet and flow straightener, a particle injector, a converging nozzle, a diverging nozzle, and a test sample mounted on a holder. GO_2 and the particle enter the inlet section and are accelerated to supersonic velocities as they pass through the converging and diverging nozzles. After flowing through a short constant-area section, the oxygen and the particle impact a sample made of the metal alloy being tested.

Typical supersonic impact of a 2000 μm- (0.08-in)-diameter aluminum particle results are shown in Figure 3. As the test sample temperature is increased, the susceptibility to ignition by particle impact increases.

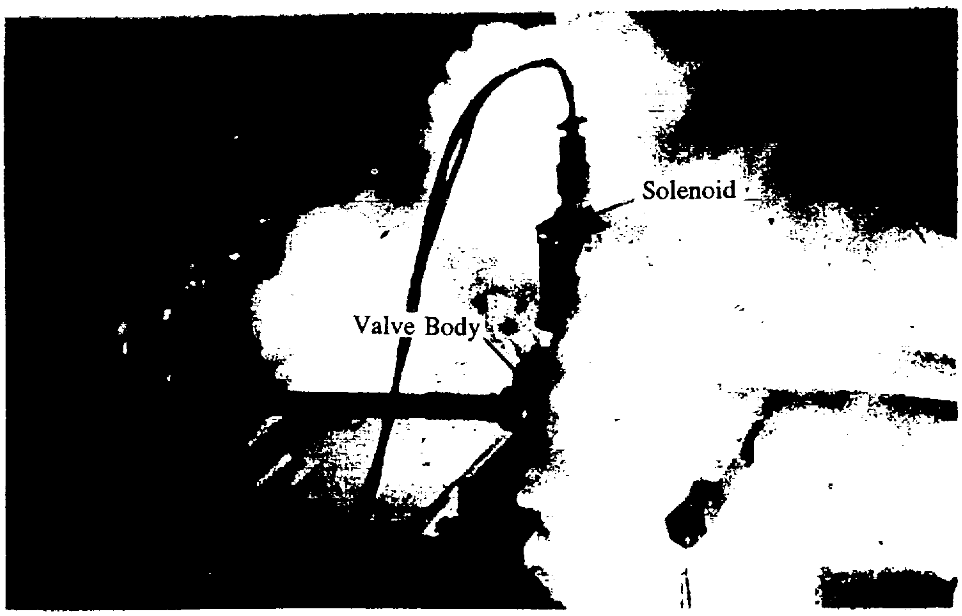


Figure 1
Valve Ignited by Metal Particles

Subsonic Particle Impact Test Method

The subsonic particle impact test chamber (Figure 4) has been described by Williams et al. [17]. It comprises a particle injector, a test sample, and a flow control orifice. Up to 5 g (0.01 lbs) of particles are injected into flowing oxygen and carried through the test chamber where they impact the test sample made from the metal alloy being tested. After impacting the sample, the oxygen and particles flow through holes on the sample periphery and are vented to the atmosphere through the flow control orifice.

Typical subsonic particle impact results are shown in Figure 5. For tests at ambient temperature with 2 g (0.004 lb) of iron powder, stainless-steel target ignitions do not occur until the particle velocity is increased to approximately 50 m/s (164 ft/s). When the temperature is increased to 360 to 450 K (188.6 to 350 °F), ignitions occur at 25 m/s (82 ft/s).

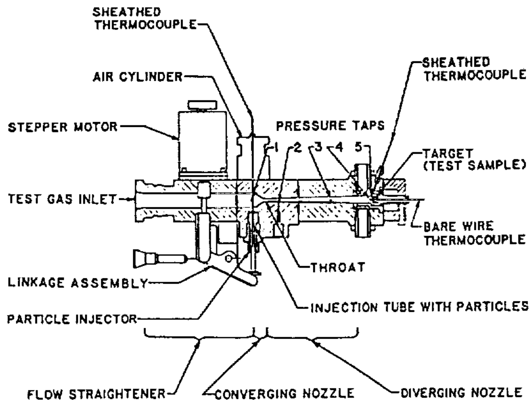


Figure 2
Supersonic Particle Impact Test Chamber

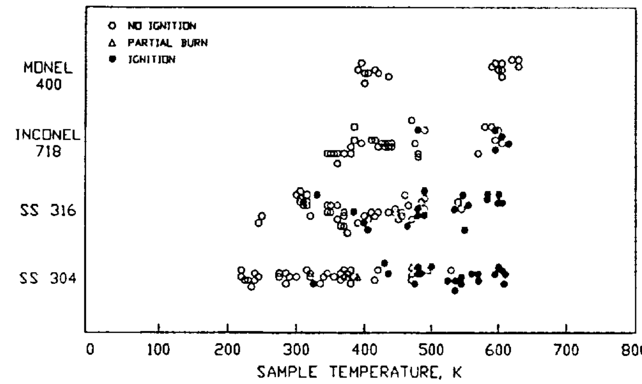


Figure 3
Typical Supersonic Particle Impact Results

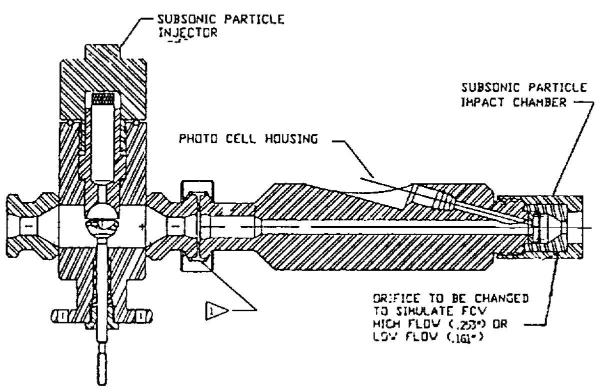


Figure 4
Subsonic Particle Impact Test Chamber

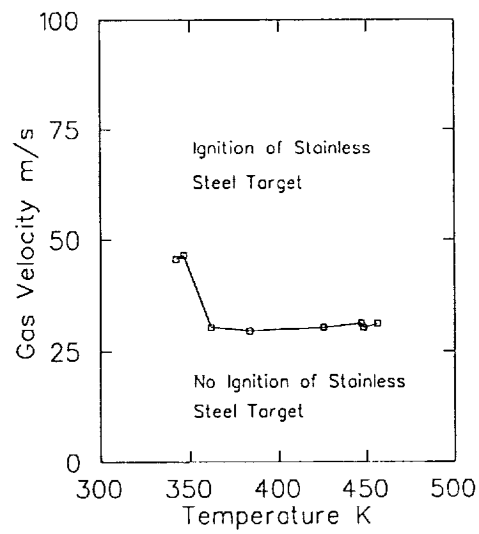


Figure 5
Typical Subsonic Particle Impact Results

The supersonic and subsonic particle impact test methods are not patented, but are available for commercial use through NASA's technology transfer program.

FRICTIONAL HEATING TEST METHOD

When mechanical components of a system rub together, heat is generated by friction. This frictionally generated heat has been identified as the cause of many fires in oxygen-enriched environments. For example, the rotating machinery shown in Figure 6 ignited and burned when a rotating part rubbed against the housing as the result of a bearing failure. Additionally, bearings in the space shuttle main engine high-pressure oxygen turbopump have ignited because of frictional heating when they failed during off-limit tests. The frictional heating test method was developed to determine metal alloys' susceptibility to frictional ignition in oxygen.

The frictional heating test apparatus shown in Figure 7 has been described by Benz and Stoltzfus [18]. It comprises an electrical drive motor and transmission assembly, a high-pressure test chamber, and a pneumatic cylinder. A rotating shaft extends through the test chamber and is connected at one end to the drive assembly

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

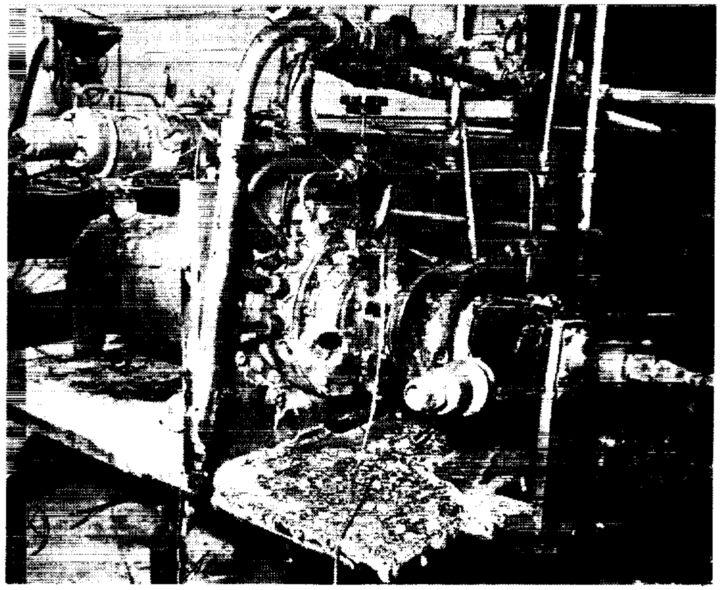


Figure 6
Rotating Machinery Ignited by Friction

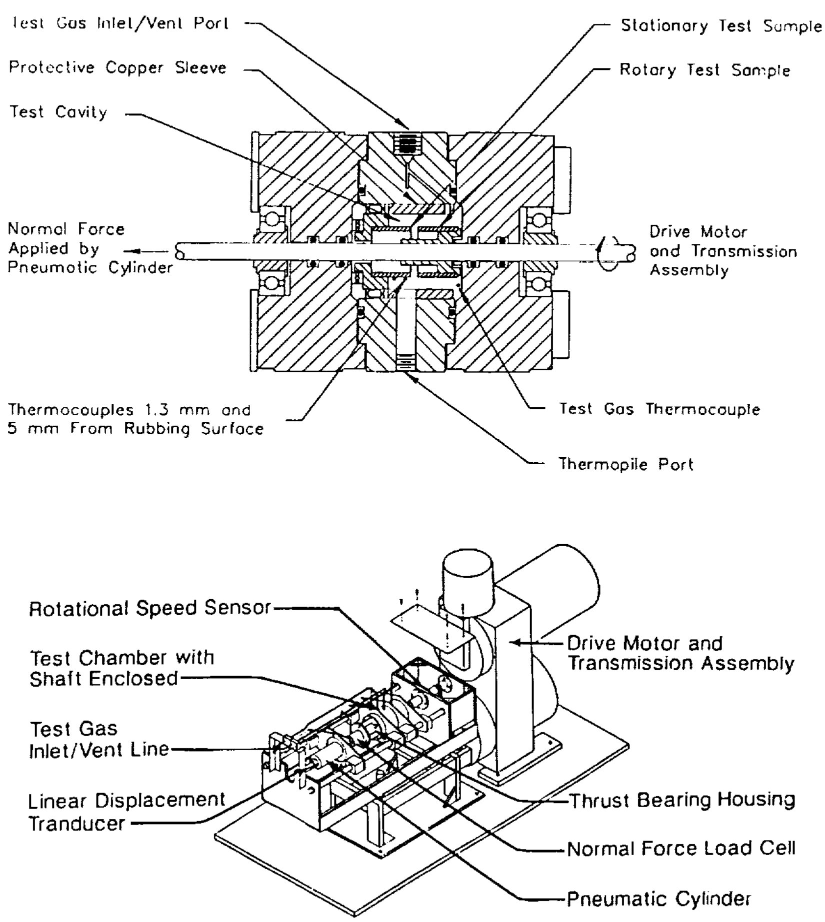


Figure 7
Frictional Heating Test Apparatus

and at the other end to the pneumatic cylinder. The rotating test specimen is mounted on the shaft, and the stationary test specimen is affixed to the test chamber. Tests up to 27.6 MPa (4000 psi) in GO_2 and 2 MPa (300 psi) in LO_2 can be conducted in this test apparatus.

Typical results of frictional heating tests are shown in Table 1. The results are presented in terms of the product of the loading pressure (P) and the rubbing velocity (v) required for ignition of the material pair.

Table 1

Results of Frictional Heating Tests in LO_2 and GO_2

MATERIAL		Pv Product ($W/m^2 \times 10^{-8}$)		Reaction	
Stationary	Rotary	GOX	LOX	GOX	LOX
Inconel 718	Inconel 718	0.96-1.18	3.85	Yes	Yes
Monel K-500	Monel K-500	1.37-1.69	4.15-4.23	Yes	No
Inco MA754	Inco MA754	3.96-4.12	3.29-4.03	No	No
Haynes 214	Haynes 214	3.05	3.79-4.09	Yes	No
AMS 6278	AMS 6278	1.82	3.34	Yes	Yes
Al 6061T6	Al 6061T6	0.48	-	Yes	---
Al 2219	Al 2219	-	1.72	---	Yes
Monel K-500	Kel-F	-	0.23-0.45[a]	---	No
Monel K-500	Vespel SP21	-	1.52	---	Yes

[a]Sample deformation before ignition

The LO_2 and GO_2 frictional heating test methods are patented and are available for licensing by NASA.

COEFFICIENT-OF-FRICTION TEST METHOD

While frictional ignition data are available using the previously described test method, a need remains to understand the tribological behavior of metals in oxygen. A test apparatus is currently under development at WSTF for this purpose (see Figure 8).

The coefficient-of-friction test method comprises an electrical drive motor and transmission assembly, a pin-on-disk frictional contact device, and a high-pressure chamber. The drive motor and transmission assembly is similar to the one used in the frictional heating test. The pin-on-disk frictional contact device comprises a cylindrical test pin with a spherical end, a disk, two shafts that penetrate the chamber, and a load arm. The pin is mounted on the load arm, which is connected to the first shaft. A measured torque applied to this shaft by means of a pneumatic rotary actuator produces the normal contact force between the pin and the 5.1-cm- (2.0-in)-diameter test disk. The disk is mounted on the second shaft, which is connected to the drive motor assembly. As the pin contacts the rotating test disk, a frictional load is detected by a calibrated strain gauge mounted on the load arm. A 5.1-cm- (2.0-in)-diameter viewport in the chamber wall allows video coverage of the test. Tests at pressures up to 27.6 MPa (4000 psi) in GO_2 and 2 MPa (300 psi) in LO_2 can be conducted in this test system. Checkout tests are being conducted and a patent is pending on the coefficient-of-friction test method. It will be available for licensing once the patent is approved.

PROMOTED COMBUSTION TEST METHOD

To determine if an alloy can be safely used in an oxygen system, its flammability must be known. If it is not flammable in a configuration similar to that in which it is intended for use, it can be used safely. The promoted

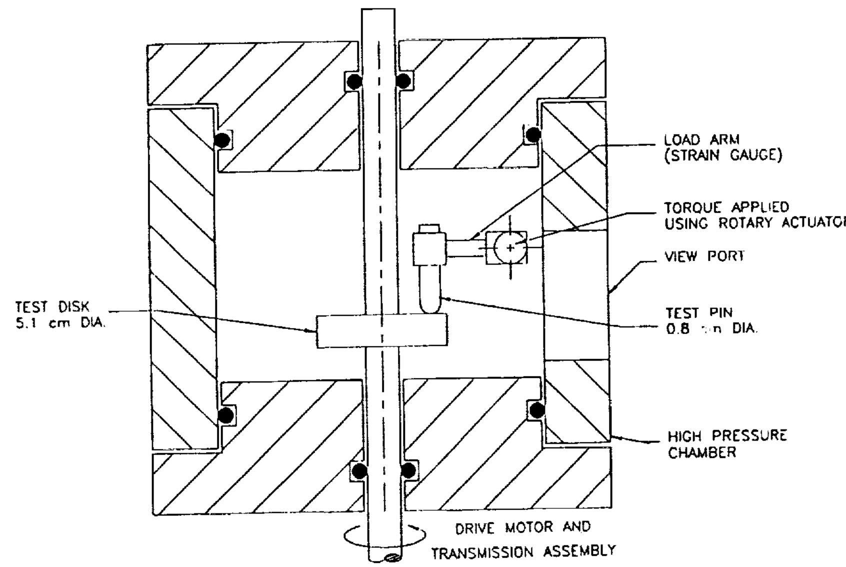


Figure 8
Coefficient-of-Friction Test Apparatus

combustion test measures flammability in terms of the minimum pressure required to support complete combustion of an alloy. This test method has been adopted by NASA as a standard for the selection of materials to be used in spacecraft. It is also being prepared as a standard by the American Society of Testing and Materials (ASTM) to determine the combustion behavior of metallic materials. The promoted combustion test apparatus shown in Figure 9 has been described by Stoltzfus et al. [19]. It comprises a cylindrical chamber, a copper liner and baseplate to protect the chamber from burning metal, and a sample mounting device. The chamber can be pressurized to 68.9 MPa (10,000 psia) and has four 5.1-cm- (2.0-in)-diameter viewports for sample observation and video recording. The test sample is held at the top in the sample mounting device and is ignited at the bottom by an aluminum or magnesium promotor. The promoter is ignited by an electrically heated wire.

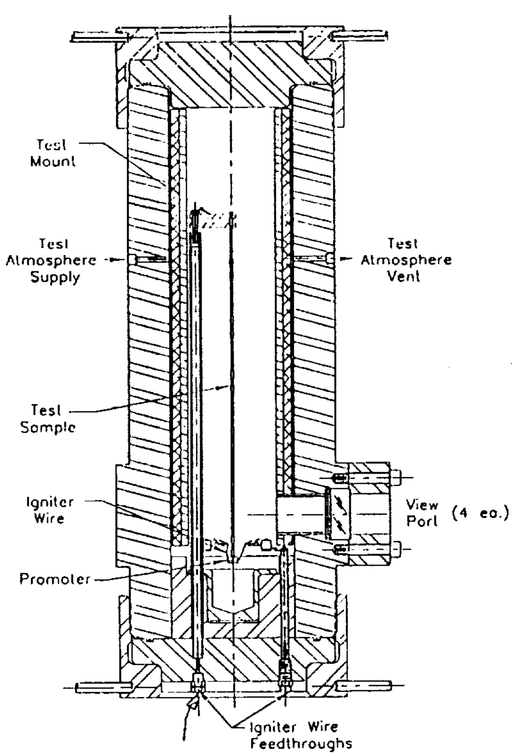


Figure 9
Promoted Combustion Test Apparatus

Typical results of promoted combustion tests on several commonly used alloys are shown in Table 2. The table presents the minimum oxygen pressure required for complete combustion of 0.32-cm- (0.125-in)-diameter metal rods (threshold pressure). Nickel, copper, and their alloys are generally the least flammable. Iron-base alloys tend to be more flammable than the nickel- and copper-base alloys, but less flammable than the aluminum- and titanium-base alloys.

The promoted combustion test apparatus is patented and is available for licensing by NASA.

Table 2
Typical Promoted Combustion Test Results

Material	Threshold Pressure MPa	(psia)
Monel K-500	68.9	>10,000
Inconel MA754	68.9	>10,000
Haynes 214	68.9	>10,000
Monel 400	68.9	>10,000
Brass 360 CDA	68.9	>10,000
Nickel 200	55.2	>8,000
Inconel 600	20.7	3,000
Inconel 625	20.7	3,000
Inconel 718	6.9	1,000
304 SS	6.9	1,000
Aluminum 6061	0.68	100
Aluminum 99.9%	0.17	25
Ti-6Al-4V	0.007	1

APPLICATION OF TEST METHODS

Johnson Space Center, Kennedy Space Center, Marshall Space Flight Center, and Lewis Research Center have used all these test methods to evaluate metals for use in oxygen systems. The results have been used to select metal alloys for the shuttle main propulsion system (MPS) oxygen flow control valves, the shuttle MPS high pressure oxygen turbopump bearings, advanced propulsion systems for the Space Exploration Initiative (SEI), and ground support equipment. The Langley Research Center has used particle impact data to select materials to use in the transpiration-cooled nozzle design for a wind tunnel.

The Department of Defense has used the test methods to determine the relative compatibility of aluminum alloys that may be used in cryogenic propellant tankage in future advanced launch systems. The data obtained have also been used to determine ignition and combustion hazards on aircraft main, backup, and emergency oxygen supply systems. Finally, the ground carts that are used to resupply aircraft oxygen systems on the flight line are being evaluated and new designs are being considered, based on the data from these test methods.

ASTM sponsored a test program that used the particle impact, frictional heating, and promoted combustion test methods to test alloys used in industrial oxygen systems. ASTM has published over 21 papers in their Standard Technical Publications containing data from these test methods [20-24]. ASTM has also prepared a training course entitled "Controlling Fire Hazards in Oxygen Systems" that makes extensive use of data from these methods. Additionally, several commercial companies have used these test methods to make material selections for applications ranging from aircraft turbine engines to compressor seal design.

These test methods are not only used by engineers for designing safe oxygen systems but they are also used for scientific research in the field of metals combustion. They have been used as tools for understanding the theories of ignition and combustion of metals and alloys [25-27]. It is anticipated that this understanding will aid in the development of new burn-resistant structural alloys for use in oxygen environments.

In summary, fire hazard in oxygen systems is a serious problem. However, a solution that minimizes this hazard exists using data from the test methods described in this paper. These methods have been used extensively in the past and are available for commercial use in the future.

REFERENCES

1) NFPA 53M. "Fire Hazards in Oxygen-Enriched Atmospheres." 1990 Edition in ASTM Standards Technology Training "Fire Hazards in Oxygen Systems." Edited by B. L. Werley, 1991.

2) de Richmond, A. L. "Laser Resistant Endotracheal Tubes--Protection Against Oxygen-Enriched Airway Fires During Surgery?" Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Fifth Volume, ASTM STP 1111. Edited by J. M. Stoltzfus and K. McIlroy, American Society for Testing and Materials, Philadelphia, 1991.

3) Sidebotham, G. W., G. L. Wolf, J. Stern, and R. Aftel. "Endotracheal Tube Fires: A Flame Spread Phenomenon." Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Fifth Volume, ASTM STP 1111. Edited by J. M. Stoltzfus and K. McIlroy, American Society for Testing and Materials, Philadelphia, 1991.

4) Bruley, M. E., and C. Lavanchy. "Oxygen-Enriched Fires During Surgery of the Head and Neck." Symposium on Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Fourth Volume, ASTM STP 1040. Edited by J. M. Stoltzfus, F. J. Benz, and J. S. Stradling, American Society for Testing and Materials, Philadelphia, 1989.

5) Simpson, J. I., G. L. Wolf, and G. A. Schiff. "The Oxidant O_2 (Helium) Index Flammability of Endotracheal Tubes." Symposium on Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Fourth Volume, ASTM STP 1040. Edited by J. M. Stoltzfus, F. J. Benz, and J. S. Stradling, American Society for Testing and Materials, Philadelphia, 1989.

6) Dicker, D. W. G. and R. K. Wharton. "A Review of Incidents Involving the Use of High-Pressure Oxygen from 1982 to 1985 in Great Britain" Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Third Volume, ASTM STP 986. Edited by D. W. Schroll, American Society for Testing and Materials, Philadelphia, 1988.

7) Barter, S. A. and L. W. Hillen. "Oxygen Fires, Materials Compatibility and System Contaminants" Symposium on Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Fourth Volume, ASTM STP 1040. Edited by J. M. Stoltzfus, F. J. Benz, and J. S. Stradling, American Society for Testing and Materials, Philadelphia, 1989.

8) Newton, B. E., R. K. Langford, and G. R. Meyer. "Promoted Ignition of Oxygen Regulators." Symposium on Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Fourth Volume, ASTM STP 1040. Edited by J. M. Stoltzfus, F. J. Benz, and J. S. Stradling, American Society for Testing and Materials, Philadelphia, 1989.

9) Bond, A.C., H. O. Pohl, N. H. Chaffee, W. W. Guy, C. S. Alton, R. L. Johnston, W. J. Castner, and J. S. Stradling. "Design Guide for High Pressure Oxygen Systems", NASA Reference Publication 1113.

10) Wegener, W., C. Binder, P. Hengstenberg, K. P. Herrmann, and D. Weinert. "Tests to Evaluate the Suitability of Materials for Oxygen Service." Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Third Volume, ASTM STP 986. Edited by D. W. Schroll, American Society for Testing and Materials, Philadelphia, 1988.

11) Lockhart, B. J., M. D. Hampton, and C. Bryan. "The Oxygen Sensitivity/Compatibility Ranking of Several Materials by Different Test Methods." Symposium on Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Fourth Volume, ASTM STP 1040. Edited by J. Stoltzfus, F. J. Benz, and J. S. Stradling, American Society for Testing and Materials, Philadelphia, 1989.

12) Lapin, A. "Liquid and Gaseous Oxygen Safety Review." Vol. I-IV, NASA-CR-120922, APCI TM184, NASA Lewis Research Center, Cleveland, OH, June 1972.

13) Clark, A. F. and T. G. Hurst. "A Review of the Compatibility of Structural Materials with Oxygen." AIAA Journal, Vol. 12, No. 4, April 1974, pages 441-454.

14) Wegener, W. "Investigations on the Safe Flow Velocity to be Admitted for Oxygen in Steel Pipe Lines." Stahl and Eisen, Vol. 84, No. 8, 1964, pages 469-475.

15) Williams, R. E. "The Combustion of Laser Ignited Zirconium." Final report NSG 518, NASA, Washington, DC, 1967.

16) Benz, F. J., R. E. Williams, and D. A. Armstrong. "Ignition of Metals and Alloys by High-Velocity Particles." Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Second Volume, ASTM STP 910. Edited by M. A. Benning, American Society for Testing and Materials, Philadelphia, 1986.

17) Williams, R. E., F. J. Benz, and K. McIlroy. "Ignition of Steel Alloys by Impact of Low-Velocity Iron/Iron Particles in Gaseous Oxygen." Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Third Volume, ASTM STP 986. Edited by D. W. Schroll, American Society for Testing and Materials, Philadelphia, 1988.

18) Benz, F. J., and J. M. Stoltzfus. "Ignition of Metals and Alloys in Gaseous Oxygen by Frictional Heating." Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Second Volume, ASTM STP 910. Edited by M. A. Benning, American Society for Testing and Materials, Philadelphia, 1986.

19) Stoltzfus, J. M., J. M. Homa, R. E. Williams, and F. J. Benz. "ASTM Committee G-4 Metals Flammability Test Program: Data and Discussion." Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Third Volume, ASTM STP 986. Edited by D. W. Schroll, American Society for Testing and Materials, Philadelphia, 1988.

20) Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: First Volume, ASTM STP 812. Edited by B. L. Werley, American Society for Testing and Materials, Philadelphia, 1983.

21) Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Second Volume, ASTM STP 910. Edited by M. A. Benning, American Society for Testing and Materials, Philadelphia, 1986.

22) Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Third Volume, ASTM STP 986. Edited by D. W. Schroll, American Society for Testing and Materials, Philadelphia, 1988.

23) Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Fourth Volume, ASTM STP 1040. Edited by J. M. Stoltzfus, F. J. Benz, and J. S. Stradling, American Society for Testing and Materials, Philadelphia, 1989.

24) Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Fifth Volume, ASTM STP 1111. Edited by J. M. Stoltzfus and K. McIlroy, American Society for Testing and Materials, Philadelphia, 1991.

25) Sato, J. and T. Hirano. "Fire Spread Limits Along Metal Pieces in Oxygen" Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Third Volume, ASTM STP 986. Edited by D. W. Schroll, American Society for Testing and Materials, Philadelphia, 1988.

26) Steinberg, T. A. and F. J. Benz. "Iron Combustion in Microgravity" Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Fifth Volume, ASTM STP 1111. Edited by J. M. Stoltzfus and K. McIlroy, American Society for Testing and Materials, Philadelphia, 1991.

27) Sircar, S., H. Gabel, J. M. Stoltzfus, and F. J. Benz. "The Analysis of Metals Combustion Using a Real-Time Gravimetric Technique" Flammability and Sensitivity of Materials in Oxygen-Enriched Atmospheres: Fifth Volume, ASTM STP 1111. Edited by J. M. Stoltzfus and K. McIlroy, American Society for Testing and Materials, Philadelphia, 1991.

N92-22446

A MAJOR ADVANCE IN POWDER METALLURGY

Brian E. Williams
Jacob J. Stiglich Jr.
Richard B. Kaplan
Robert H. Tuffias

Ultramet
12173 Montague Street
Pacoima, CA 91331

ABSTRACT

Under SBIR funding from the Army Materials Technology Laboratory, Ultramet has developed a process which promises to significantly increase the mechanical properties of powder metallurgy (PM) parts. Current PM technology utilizes mixed powders of various constituents prior to compaction. The homogeneity and flaw distribution in PM parts depends on the uniformity of mixing and the maintenance of uniformity during compaction. Conventional PM fabrication processes typically result in non-uniform distribution of the matrix, flaw generation due to particle-particle contact when one of the constituents is a brittle material, and grain growth caused by high-temperature, long-duration compaction processes. Additionally, a significant amount of matrix material is usually necessary to fill voids and create 100% dense parts. In Ultramet's process, each individual particle is coated with the matrix material, and compaction is performed by solid state processing. In the AMTL program, Ultramet coated 12-micron tungsten particles with approximately 5 wt% nickel/iron. After compaction, flexure strengths were measured 50% higher than those achieved in conventional liquid phase sintered parts (10 wt% Ni/Fe). This paper presents further results of the AMTL program and discusses other material combinations.

INTRODUCTION

Standard powder metallurgy techniques rely on the physical mixing of powder constituents in order to provide a homogeneous composite or alloy. Many approaches have been developed to permit better mixing, often focusing on reducing particle size: the smaller the size, the lesser the adverse effects of any isolated areas of nonhomogeneity. Small particles, however, lead to other problems such as increased surface area, which promotes greater reactivity with any embrittling impurities that are present. Also, fabrication costs are higher due to the difficulties encountered in handling fine particulates.

Ultramet's Army-sponsored research takes a different approach to promoting homogeneity. Through the use of chemical vapor deposition (CVD), Ultramet has demonstrated the capability of coating individual tungsten particles (5 to 20 microns in diameter) with varying levels of nickel, iron, cobalt, and combinations of the three. This method of "premixing" the composite constituents, exceeds by far the current capabilities of powder metallurgy, as the distribution of the matrix material is accomplished at the individual powder particle level. The ductility of metallic powder is ultimately enhanced, a highly desirable property for many applications.

BACKGROUND

Chemical Vapor Deposition (CVD)

CVD is a coating method that utilizes the decomposition of a gaseous precursor, flowed over or through a heated substrate, with subsequent condensation from the vapor state to form a solid deposit. The CVD process is an extremely versatile and relatively inexpensive method of molecular forming. Its benefits include the potential to produce deposits of controlled density, thickness, and composition, with extremely low impurity levels.

CVD has been successfully utilized for coating particles, including the production of ultrapure tungsten and niobium spheroids for metallurgical purposes and cladding of nuclear fuel (UO_2) particles. As such, it represents a proven approach with a large payoff. CVD has been used for the deposition of some 400 species, the processes for many of which were pioneered by Ultramet personnel.

The CVD process itself promotes greater purity, particularly for powder coating, as in situ gettering of contaminants can be performed in order to purify both the as-received (uncoated) powder and the resultant coatings. Also, the ability of CVD to produce multilayered or alloyed coatings allows for precise control of the desired physical and chemical characteristics of the final fabricated shape.

Fluidized-Bed CVD

In order to uniformly coat fine particulates, fluidized-bed technology has been used. A schematic of a typical fluidized-bed CVD reactor, used for coating particles, is shown in Figure 1. The reactant gas stream is united with the powder fluidization gas just prior to entering the reactor. The metal precursor then preferentially decomposes on the surface of suspended powder particles entering the heated reaction zone, which act as nucleation sites for continuous film growth.

The gas flow in the entrance orifice must be greater than the terminal velocity of the particles so that particles do not fall down the orifice, and the gas velocity in the parallel section above the orifice must be less than the terminal velocity of the particles so that they are not blown out of the coater. In order to fluidize the particles, the gas velocity in the parallel section must be greater than the minimum fluidization velocity, u_{mf}, which is given by

$$u_{mf} = \frac{d_p^2(\rho_s - \rho_g)G}{1650\mu} \quad \left(\text{for } \frac{d u_o \rho_g}{\mu} < 20\right) \tag{1}$$

where d_p is the particle diameter, ρ_s is the density of solids, ρ_g is the fluid density, G is the acceleration of gravity, μ is the gas viscosity, and u_o is the superficial gas velocity.

The reactor shown in Figure 1 was designed to remove powder fines in order to produce an extremely narrow particle size distribution. The lighter fines are carried by the fluidizing gas stream into the disentrainment section, where they settle. Part of the proposed effort is aimed at developing scaling relations to translate the process to different particle sizes and densities.

Ultramet has developed a CVD fluidized-bed reactor capable of both fluidizing and coating metallic and ceramic particles as small as 5 μm in diameter with a large number of different metal and ceramic materials. Efficient fluidization was previously limited to >15-μm particle diameters. The reduced size now available increases the driving force in sintering, since the sintering rate is roughly proportional to the inverse of the particle size. The resultant improved densification is a result of several concurrent processes. Smaller particle size leads to an increase in the energy associated with solid/pore interfacial areas, increasing the driving force available for compaction. Also, the greater interparticle contact provides more paths for volume diffusion or material transport, and the greater surface area allows for an increase in grain boundary/surface diffusion.

Powder particles are coated free of agglomeration, while the concentration and thickness of the deposited coating can be easily controlled. Standard powder metallurgy techniques involving powder mixing often lead to increased porosity and elemental heterogeneity. The use of prealloyed (uncoated) matrix powder particles eliminates the problem of heterogeneity, but these powders are very hard and compaction is difficult. At normal compaction pressures, the plastic deformation needed to obtain the amount of interfacial contact necessary for effective diffusion is difficult, and porosity results. The use of composite (coated) powders eliminates heterogeneity, allowing for a level of compaction comparable to or better than that available from elemental powders.

One of the primary advantages of CVD over most plating methods is the extremely high level of purity which may be obtained in the deposit. The majority of impurities in a coated powder batch is due to contamination in the as-received, uncoated substrate powder itself. In the case of deposition on tungsten powder, Ultramet has shown that light-element impurity levels (especially carbon and oxygen) may be substantially reduced through heat treatment, hydrogen reduction, and/or controlled water vapor treatment of the as-received tungsten powder. Carbon levels were reduced by a factor of twelve from the as-received, uncoated powder to the coated, treated powder, and oxygen levels were reduced by a factor of four. Reduction of embrittling impurities leads to a substantial increase in mechanical properties.

EXPERIMENTAL APPROACH

Ultramet's Army-sponsored research initially developed the capability to coat 12-micron tungsten particles with nickel, iron, and nickel/iron and nickel/cobalt mixtures via fluidized-bed chemical vapor deposition. Test specimens were then fabricated by densifying the coated tungsten particles via various consolidation methods; their microstructures were characterized; and their mechanical properties were measured at both quasistatic and elevated strain rates. Consolidation and testing focused on a (nominal) composition of 95 wt% tungsten/5 wt% nickel-iron, with the Ni:Fe ratio being 70:30 (overall composition 95W:3.5Ni:1.5Fe).

Fabrication/Consolidation Technologies

Three fabrication/consolidation technologies were evaluated during the course of this work: liquid phase sintering (LPS), the Ceracon process, and hot isostatic pressing (HIP).

LPS was performed by AMTL as a baseline consolidation technology. The Ceracon process (Ceracon Inc., Sacramento, CA), a derivative of HIP involving dynamic force that induces compaction by both physical and activated sintering means, yielded encouraging results. HIP consolidation was performed by IMT (Portland, OR) on coated powders contained in evacuated stainless steel cans. Consolidation conditions (pressures, temperatures, times) for these methods are shown in Table I. A comparison between the microstructures obtained by Ceracon consolidation of Ultramet coated powder, and standard liquid phase sintering of mixed powders, is shown in Figures 2A and 2B. The uniformity of matrix dispersion in the Ultramet/Ceracon material is clearly evident.

Mechanical Test Specimen Preparation

Three-point flexure testing was chosen to evaluate both strength and ductility by measuring midspan deflection and maximum flexural stress. Specimen dimensions were determined from the consolidated billet size available; therefore, the data shown should be used only to draw general conclusions about the various materials tested. Specimen dimensions are listed in Table I for all samples tested.

Three forms of consolidated W/Ni-Fe composite materials were tested: a commercial LPS material, a HIP material, and a Ceracon-consolidated material. The HIP and Ceracon processes utilized W:3.5Ni:1.5Fe powder that was fabricated by CVD at Ultramet. The LPS material was W:7.0Ni:3.0Fe, which was provided by AMTL in the form of an unworked, quarter-scale penetrator. The specimens were tested at ambient temperature using a cross-head speed of 0.25 mm/min (0.010"/min).

Hopkinson bar compression testing was performed at the University of California at San Diego (UCSD). Hopkinson bar specimens were 3.8-5.1 mm (0.150-0.200") diameter x 4.0 mm (0.156") long (the dimensions varying according to the size of the billet available).

RESULTS AND DISCUSSION

Hopkinson bar compression test data, produced at UCSD, showed true strains of approximately 28% for 3900/sec and 4000/sec strain rates, and 76% for a 5000/sec strain rate. Engineering strains were $\approx$22% for the lower strain rate and 53% for the higher strain rate, indicating the presence of some elastic recovery in these composites.

Table I shows the results of flexure testing, including specimen size and composition, processing conditions, and midspan deflection (an indication of ductility). Figures 3A-3D show the results of metallographic analysis performed on HIP, LPS, and Ceracon consolidated materials. The consolidation conditions required to achieve 100% theoretical density via HIP were 1185°C (2165°F), 172 MPa (25 ksi), and 4.0 hours. The resultant microstructure (Figure 3A) exhibited substantial contact between tungsten particles and poor matrix distribution; no attempt was made to optimize the HIP cycle. The 5% matrix HIP material exhibited ultimate stresses that were comparable to the 10% matrix LPS material (Figure 3B); however, the LPS material exhibited a greater midspan deflection. The processing conditions for the LPS material were much more complex, involving oxide reduction treatment before sintering, sintering at 1760°C (3200°F), and by a vacuum degas treatment. Sintering time is typically 30 min for a simple quarter-scale penetrator shape.

The most favorable combination of strength and deflection was exhibited by the 6% matrix material consolidated through the Ceracon process (Figure 3C). Consolidation conditions were 1235°C (2255°F), 1379 MPa (200 ksi), and 30 sec. This 6% matrix material exhibited strengths and midspan deflections over 50% greater than those of the 10% matrix commercial material. Figure 4 is a photograph of a W:4.5Ni:1.5Fe flexure test bar, showing an unusual degree of bending for such a tungsten-heavy material.

The differences in mechanical behavior may be explained by the homogeneity of matrix distribution, which was a primary goal of this work: to demonstrate that improving the matrix dispersion over standard PM techniques will improve physical properties as well. The SEM micrographs of these materials (Figures 3A, 3B, and 3C) support this hypothesis.

The Ultramet CVD/HIP material (Figure 3A) exhibited an almost entirely intergranular fracture mode. The sharp-edged, faceted fracture surfaces resulted from the brittle fracture behavior that is expected at the interface of tungsten grains in direct contact. The matrix material, initially designed to provide an interface between tungsten grains, was extruded into triple point locations between grains. These high matrix content areas are clearly evident in the micrographs. In addition, a portion of the intergranular fracture most likely involved the breaking up of the several tungsten single crystals that made up each tungsten grain, prior to CVD coating.

The commercial LPS material (Figure 3B) exhibited fracture in virtually a single plane of tungsten particles. The flat areas shown on individual tungsten grains were interfaces between tungsten particles with no matrix material present. Figure 3D, meanwhile, clearly shows the bright tungsten flat areas on individual tungsten grains, with no sign of the Ni-Fe matrix material. The result was again predominantly intergranular failure.

The material fabricated through the Ultramet CVD/Ceracon process (Figure 3C) exhibited fracture behavior that was significantly different from the other two materials. There was evidence of three different failure modes: intergranular, intragranular, and grain pullout. The latter was different from that shown in the HIP sample, in that the void left from the removed grain was lined with Ni-Fe matrix material. The high transverse rupture strength is attributed to the low incidence of tungsten-tungsten particle contact.

COMMERCIAL APPLICATIONS

Ultramet has demonstrated the ability to coat individual powder particles by CVD, forming a true "composite" metal by the integration of a hard metal powder reinforcing phase with a ductile matrix. The expected improvements in mechanical properties are thought to be due to a nearly perfect distribution of matrix material about the individual tungsten particles in a 95 wt% tungsten composite. Table II shows the various powder/coating combinations that have been successfully obtained at Ultramet.

The combined new technologies developed in this work will mitigate or remove the present barriers to improving powder metallurgy component fabrication. The result will be components with dependably better properties and narrower, more predictable statistical property distributions. The size and weight of load-bearing sections of components can be reduced without giving up strength, or strength can be increased without increasing size and weight. A narrow, more predictable statistical variation in mechanical properties provides more confidence in designing structures utilizing such new materials.

The military applications of this technology are obvious. The DARPA armor/antiarmor initiative has identified advanced kinetic energy weapons as an area where major improvements in U.S. capabilities are needed. The application of new, innovative powder preparation and consolidation technologies will reduce costs and provide the performance improvements necessary to penetrate present and future armor.

Improved ductility and densification of materials for penetrator devices is only one of the possible applications of this technology. In addition to ordnance applications, tungsten heavy metal alloys have utility in neutron shielding. Better mechanical properties and homogeneity may allow improvement in the performance/weight consideration for such applications. Graded seals and conductive ceramics are also areas of interest.

Ultramet's powder coating technology has many more subtle, but potentially very important, benefits to both government and commercial interests. Powder metallurgy recently has made great strides in the fabrication of difficult-to-process alloys, including metastable structures for increased wear and corrosion resistance as well as stronger alloys to permit weight reduction. The latter is of great interest to many military systems, among them helicopters, fighter aircraft, ground vehicles, and artillery.

Prospects are also very good for making use of this technology in the fabrication of cemented carbide tools and wear parts, currently a $1 billion market. It is likely that these tools' transverse rupture strength and impact strength can be measurably improved. Ultimately, the entire area of metal/ceramic composites will be impacted by this technology. One example of such an advanced composite is ceramic-strengthened intermetallic materials for use in aerospace engines and structures.

While this effort has focused on the development of improved powders for tungsten heavy metal composites, this technology is being extended to develop a wide range of powder compositions for the metals, ceramics, and composites industries. Compositions and applications ready for immediate development include:

- Custom-coated powders for plasma spraying.
- Ceramic powders integrally coated with their corresponding sintering aids. Like the tungsten composites, this leads to reduced sintering times and temperatures, reduced grain growth, reduced contamination, and improved properties and economics.
- Matrix-coated silicon carbide (SiC), aluminum oxide (Al_2O_3), boron carbide (B_4C), cubic boron nitride (CBN), diamond, and other powders leading to improved cutting tool performance and improved plasma-sprayed wear-resistant coatings.
- Matrix- and/or interface-coated whiskers and particulates leading to injection-moldable, more cost-effective composites.
- Extension of mechanical alloying technology to include interstitial-sensitive metals such as titanium and niobium. The mating of CVD and mechanical alloying theoretically allows almost any combination of materials to be fabricated in a dispersion-strengthened composite.
- Diffusion barrier/compatibility layer coating of microspheres, powders, and particulates providing wetting and stability of fillers and reinforcements. Potential applications include stabilizing filler particle oxidation state for high temperature magnetic and electronic applications, modifying the bonding and long-term stability of SiC whiskers in titanium alloys, among others.

ACKNOWLEDGMENTS

This work was performed under contract #DAAL04-88-C-0030 with the Army Materials Technology Laboratory (Watertown, MA). The Ultramet authors would like to thank the AMTL program managers, Dr. Kenneth J. Tauer and Robert J. Dowding, for their support and assistance throughout the performance of this work, and Dr. Ramas V. Raman and Sundeep Rele at Ceracon for their expert advice and assistance in powder consolidation.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

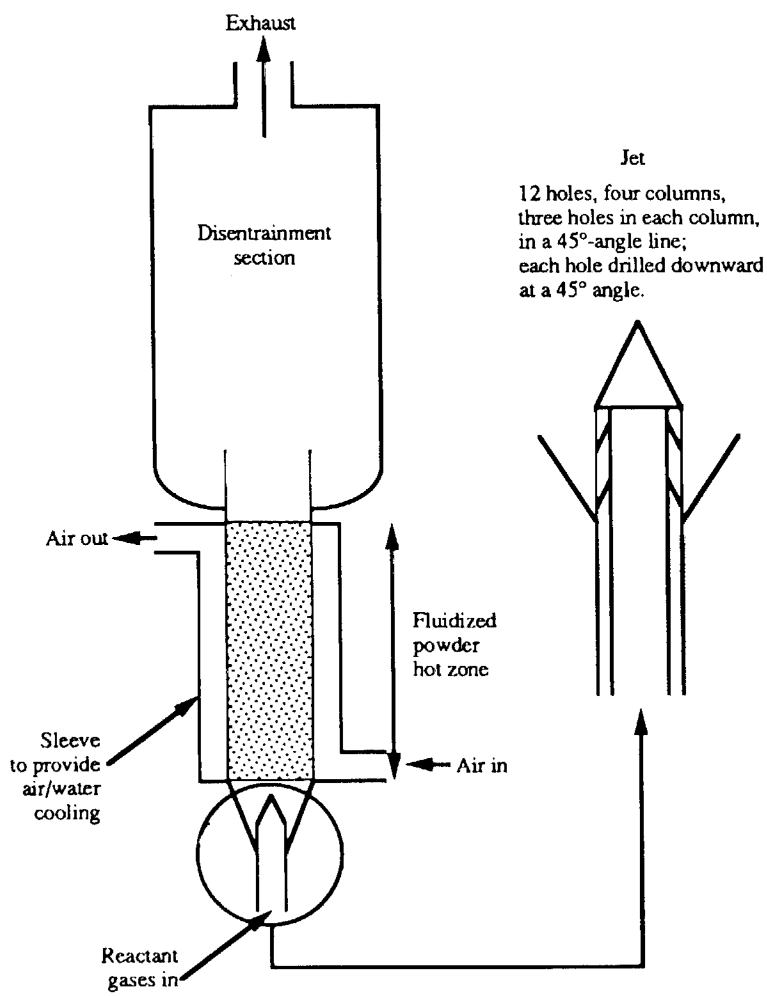


Figure 1. Schematic of fluidized-bed CVD apparatus

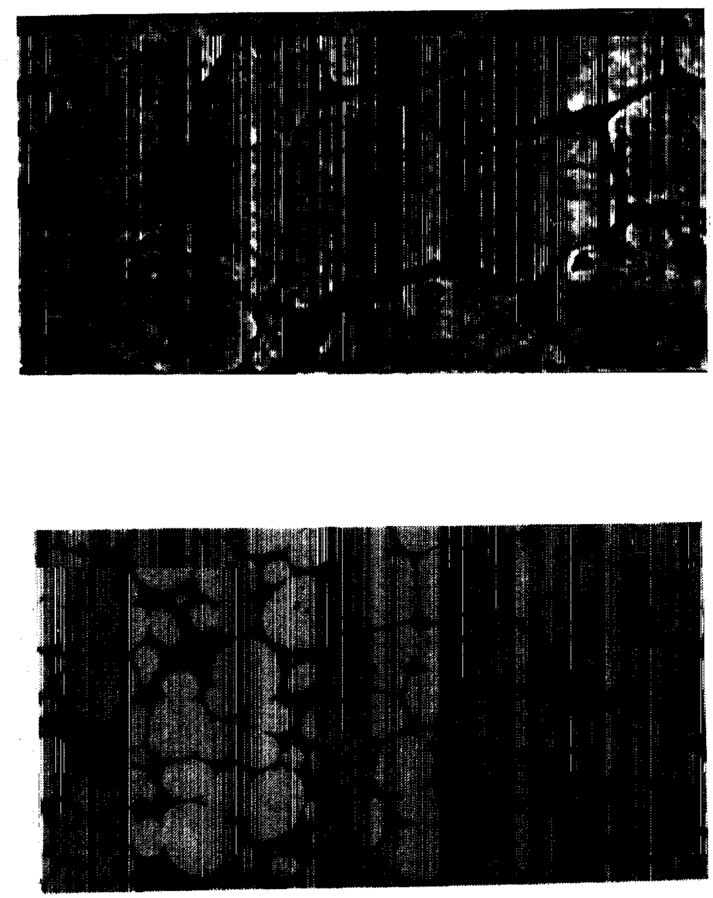


Figure 2. A (top): SEM micrograph of Ultramet W:3.5Ni:1.5Fe material consolidated by Ceracon, showing extreme uniformity of matrix dispersion (2000x)
B (bottom): SEM micrograph of AMTL W:3.2Ni:1.0Fe material consolidated by LPS, showing extreme tungsten-tungsten particle contact and nonhomogeneous matrix dispersion (400x)

Table I. Flexure Test Results (Three-Point Loading)

Length (in)	0.5625	0.5625	0.5625	0.5625	0.5625	0.5625	0.5625	0.5625	0.5625
(mm)	14.29	14.29	14.29	14.29	14.29	14.29	14.29	14.29	14.29
Width (in)	0.1903	0.1894	0.1894	0.1860	0.1890	0.1890	0.1488	0.1492	0.1494
(mm)	4.834	4.811	4.811	4.724	4.801	4.801	3.780	3.790	3.795
Depth (in)	0.1630	0.1590	0.1590	0.1490	0.1500	0.1500	0.1260	0.1255	0.1260
(mm)	4.140	4.039	4.039	3.785	3.810	3.810	3.200	3.188	3.200
Composition: %W	95.3	95.3	95.3	90.0	90.0	90.0	94.0	94.0	94.0
%Ni	3.5	3.5	3.5	7.0	7.0	7.0	4.3	4.3	4.3
%Fe	1.2	1.2	1.2	3.0	3.0	3.0	1.7	1.7	1.7
% uncoated 5-μm tungsten powder	10.0	10.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0
Consolidation	HIP	HIP	HIP	LPS	LPS	LPS	Ceracon	Ceracon	Ceracon
temp. (°F)	2165	2165	2165	2730	2730	2730	2255	2255	2255
(°C)	1185	1185	1185	1500	1500	1500	1235	1235	1235
pressure (ksi)	25	25	25	n/a	n/a	n/a	200	200	200
(MPa)	172.4	172.4	172.4				1379	1379	1379
duration (hr)	4	4	4	1	1	1	0.008	0.008	0.008
Measured Density (g/cm^3)	18.24	18.24	18.24	17.15	17.15	17.15	17.95	17.95	17.95
Theoretical Density (g/cm^3)	18.24	18.24	18.24	17.15	17.15	17.15	17.95	17.95	17.95
Percent Dense	100	100	100	100	100	100	100	100	100
Deflection (in)	0.011	0.009	0.009	0.017	0.015	0.015	0.024	0.018	0.020
(mm)	0.279	0.229	0.229	0.432	0.381	0.381	0.610	0.457	0.508
Maximum Load (lbf)	1217	1162	1169	1000	1111	1059	889	801	785
(N)	5413	5169	5200	4448	4942	4710	3954	3563	3492
Ultimate Stress (ksi)	203	205	206	204	220	210	318	288	279
(MPa)	1400	1413	1420	1407	1517	1448	2193	1986	1924

Figure 3A. SEM micrograph of W:3.5Ni:1.5Fe material consolidated by HIP (fracture surface, 2000x), showing poor matrix dispersion leading to intergranular fracture.

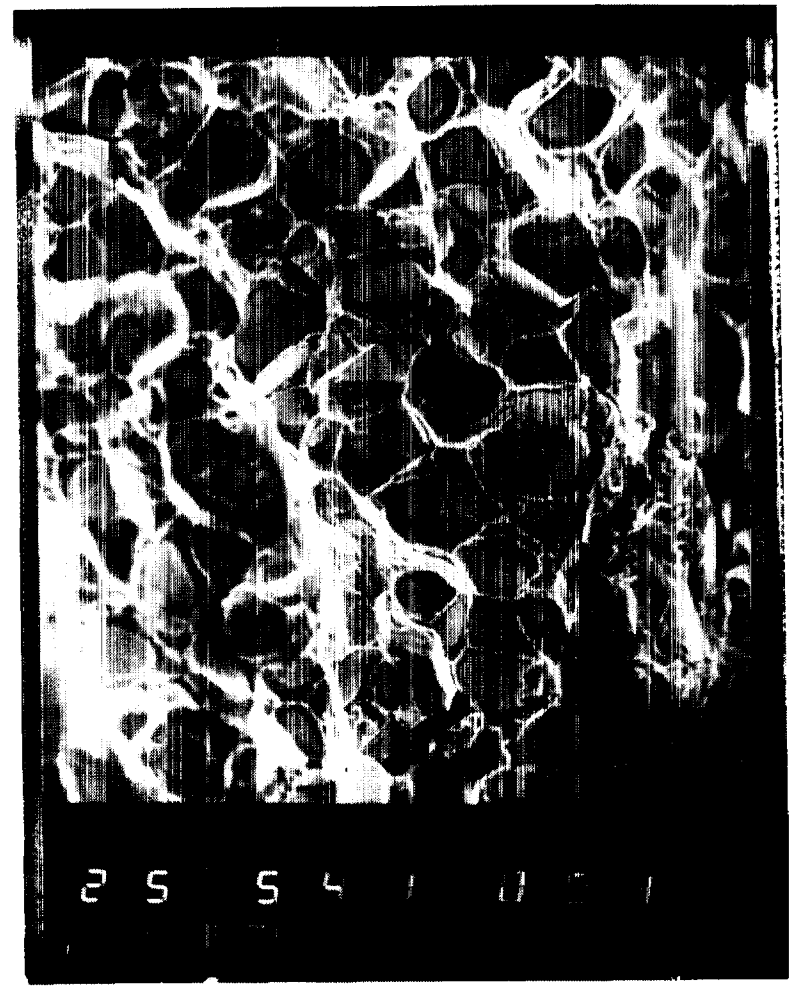


Figure 3B. SEM micrograph of W:7.0Ni:3.0Fe material consolidated by LPS (fracture surface, 540x), showing poor matrix dispersion leading to intergranular fracture originating at the flat surfaces seen on individual tungsten grains.

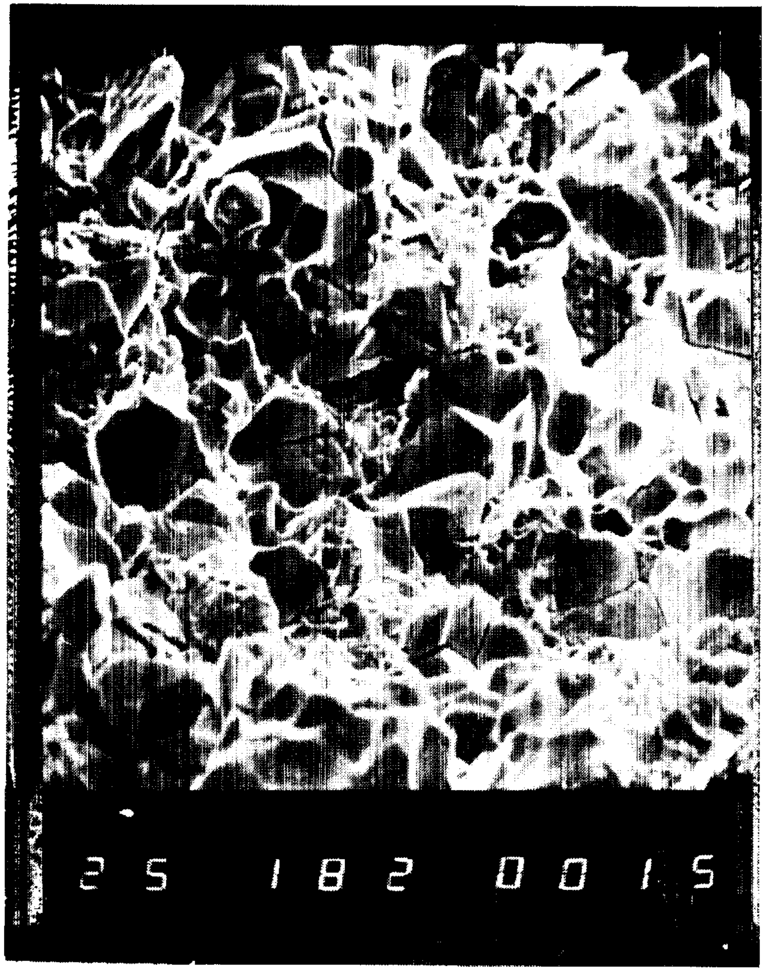


Figure 3C. SEM micrograph of W:3.5Ni:1.5Fe material consolidated by Ceracon (fracture surface, 1800x), showing significant <u>intra</u>granular failure and grain pullout in addition to intergranular fracture.

Figure 3D. SEM micrograph of W:7.0Ni:3.0Fe material consolidated by LPS (secondary [top] and backscattered [bottom] electron images, fracture surface, 2000x). The shiny flat surfaces on individual grains were areas of tungsten-tungsten particle contact prior to fracture; no sign of Ni-Fe matrix material is evident in these areas.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

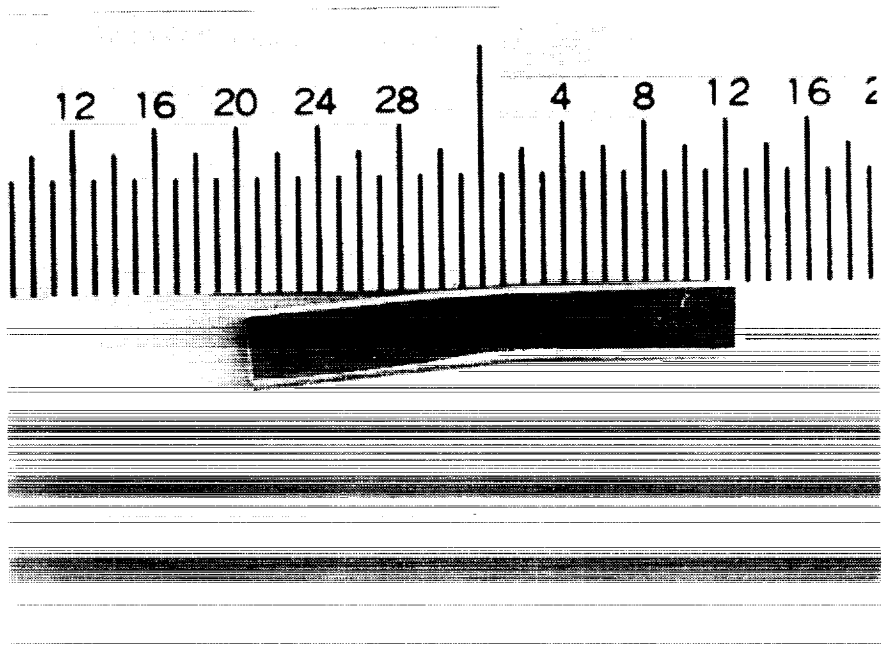


Figure 4. **W:4.5Ni:1.5Fe composite flexural specimen following three-point loading test, which was aborted just prior to expected fracture in order to show degree of bending exhibited.**

Table II. Powder Coating Programs at Ultramet

Description (coating on powder)	Application
3.5Ni:1.5Fe on 12-μm tungsten	Ordnance
5-50 wt% copper on 100-μm AlN	High conductivity composites
10-30 wt% aluminum on 5-μm TiB_2	Dispersion strengthening
TiB_2 on 5-μm aluminum	Dispersion strengthening
80 wt% tungsten on 150-μm Al_2O_3	Proprietary
10 wt% Al_2O_3 on 100-μm SiC	Ceramic composites
20 wt% titanium on 100-μm Al_2O_3	Proprietary
5 wt% iron on 100-μm WC	Cutting tools
3 wt% cobalt on 10-μm WC	Cutting tools
3 wt% iron on 20-500-μm diamond	Cutting tools
10-20 wt% hafnium and titanium on 12-μm tungsten	Ordnance

N92-22447

PERMANENT MAGNET DESIGN METHODOLOGY

Herbert A. Leupold
U.S. Army Electronics Technology and Devices Laboratory
Fort Monmouth, NJ 07703-5601

ABSTRACT

Design techniques developed at ETDL for the exploitation of high energy magnetically rigid materials such as Sm-Co and Nd-Fe-B have resulted in a revolution in kind rather than in degree in the design of a variety of electron guidance structures for ballistic and aerospace applications. Some of the salient examples are:

1) magnets for traveling wave tubes that have from one to two orders of magnitude advantage in field-to-mass ratio over conventional structures.

2) permanent magnet solenoids for high-powered klystrons that generate thousands of gauss of field uniformly over cylindrical volumes that are a meter long and one half meter in diameter.

3) sources of transverse magnetic fields of several kilogauss and parts-per-million uniformity for use in magnetic resonance imagers (MRI) for medical diagnostics, airport baggage scanners and general lab use.

4) compact, light-weight, high-field, finely tunable free electron laser magnets that require no power supply and exhibit minimal stray fields.

5) two-tesla magnetic fields generated by grapefruit-sized structures in one inch spherical cavities for optical Faraday rotators and short path traveling wave tubes.

INTRODUCTION

A perennial barrier to the application of the latest high-powered radiation sources to airborne, ballistic, and the more highly mobile surface vehicles has been the excessive mass, bulk and dependence on power packs of the electron beam focussing magnets that such sources employ. Until relatively recently, attainment of magnetic fields of several thousand gauss over large gaps or volumes depended upon bulky electro-magnets with equally cumbersome power supplies or on large masses of conventional magnet materials whose weight and bulk severely limited application to mobile devices. Many field configurations were unattainable even with combinations of extraordinarily large mass, high current, and small volume.[1,2]

With the advent of the magnetically rigid high energy product rare earth permanent magnet materials (REPM's), these difficulties became tractable and whole families of previously unattainable devices became viable. Such materials are characterized by very high remenance and coercivity. The former is a measure of a material's ability to provide large amounts of magnetic flux, while the latter is a measure of its rigidity, that is, ability to maintain magnetization in the face of strong opposing magnetic fields that arise from geometric demagnetization effects or from externally applied sources. Such fields are always present in permanent magnet devices, especially in the compact structures that are so important to the military.

Although the high energy product materials (Samarium-cobalt, neodinium-iron-boron, mischmetal, etc.) have been commercially available for almost two decades, they have not been exploited to the revolutionary extent warranted by their properties. This seems to be because force of custom causes designers to employ the new materials to improve performance of old structures rather than to use them in entirely new designs that are not practicable with the old materials. In this regard, the magnetics group at ETDL has been an exception in, that over the past decade, it has striven to formulate general design principles needed to afford these principles to obtain efficient, compact, and lightweight devices viz radars, radios, electronic warfare, fuses, magnetic resonance imagers for medical diagnostics, motors, generators and others. This work has resulted in the disclosure of over one hundred patents and the construction of several prototype models of which some of the most advanced will be discussed in this paper. These structures fall roughly into four broad classes with some overlap.

Permanent Magnet Solenoids

Permanent magnet solenoids (PMS) provide uniform fields of thousands of oersteds over considerable lengths in cylindrical structures.[3,4] Before the advent of REPM's, such fields were attainable only with electrical solenoids that were generally cumbersome, consumptive of electrical energy and dependent on electric current sources for their operation. The latter are not conveniently portable and do not lend themselves readily to providing field to small spaces with compact structures. In contrast, permanent magnet solenoids generate fields of up to slightly more than half of the remanances of the permanent magnets used (typically about 10kG for REPM's) in cylindrical spaces of arbitrary dimension without the drawbacks of electrical solenoids. The generated magnetic flux is essentially confined to the device and stray fields are minimal. This is accomplished by a configuration that maintains an equipotential everywhere on its outer surface. These configurations consist of three parts: an axially magnetized cylindrical shell which supplies the flux; iron discs that guide that flux into and out of the ends of the cylindrical cavity; and radially magnetized conical cladding which together with axially magnetized discs at the ends, and obliquely magnetized rings in the corners, confines the flux to the interior. Variations of the basic design can produce fields with gradients along the principal axis while maintaining field uniformity over any cross section.[3] Other variations can confine flux to annular ring shaped regions.[5] The PM's are useful in various electron beam devices such as traveling wave tubes, klystrons, cross-field amplifiers and gyrotrons.[6] Representative structures are shown in Figures 1-6 in which magnetization vectors are represented by thin arrows and the working field vectors by thick arrows.

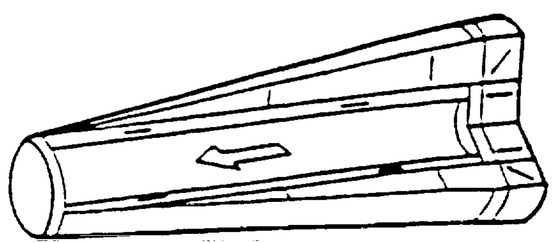


Figure 1. Single chambered Neugebauer structure. Left pole piece is taken as zero potential and outer surface everywhere is lowered to the same potential by inward-pointing cladding magnets. In this way, flux is confined to the interior.

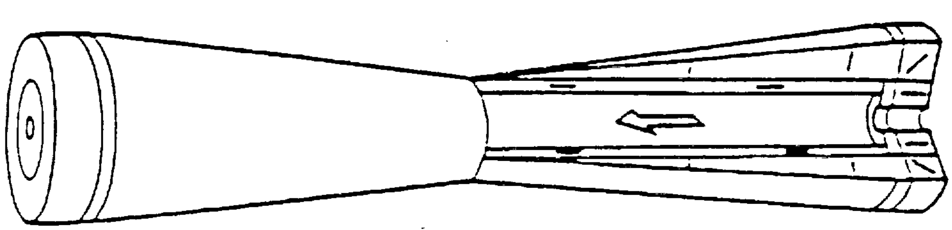


Figure 2. Because the zero potential reference has ben moved to the middle from the end, the ETDL structure pictured here has less than half the mass of the Neugebauer design of Figure 1. Solenoidal fields up to 5 kOe are easily produced in this manner, in a structure much lighter than an equivalent electric solenoid and its power supply.

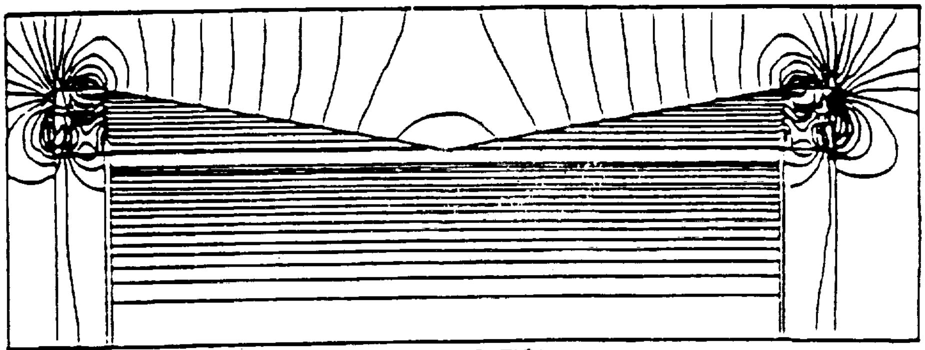


Figure 3. Flux plot of magnetic field produced by the permanent-magnet solenoid of Figure 2. Note the great uniformity over the working space. Apparent flux crowding towards the periphery is because each line represents a unit of flux in an annular ring of given thickness. Actual leakage is only a few percent and is due to imperfect cladding at the corners.

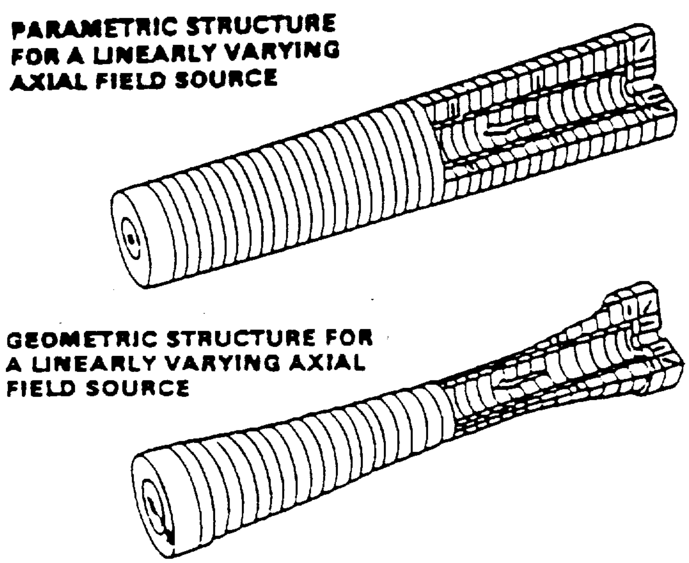


Figure 4. The ETDL permanent-magnet solenoid can be modified to produce fields that vary along the axis. Both of the pictured structures accomplish this. The parametric version controls the variation by modulation of the magnetic properties. In the geometric version, structural dimensions are modulated.

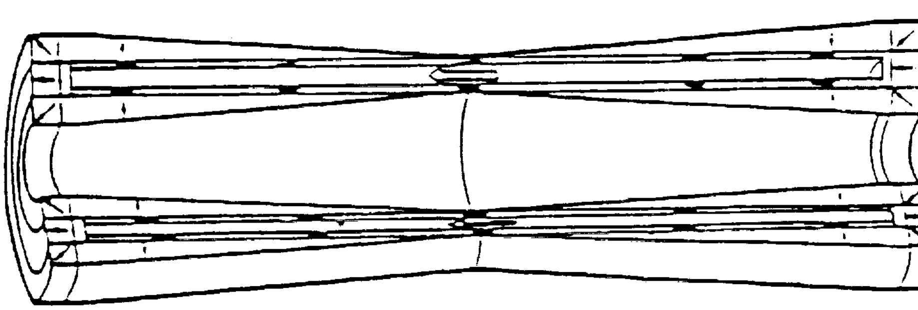


Figure 5. The field in the annular working space is supplied by the axially oriented magnetic shells. The radially oriented shells and end magnets confine the flux. Applications include hollow beam devices.

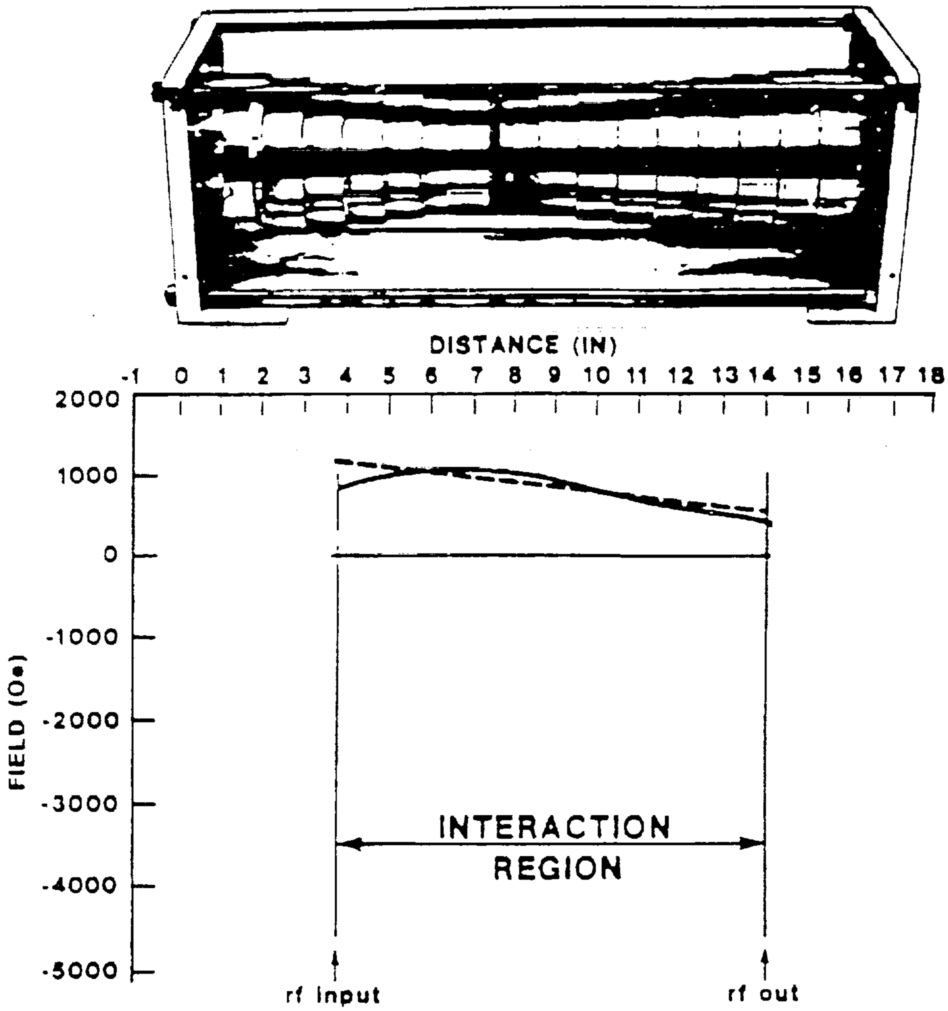


Figure 6. Axially tapered field solenoid.

Figure 6 shows an axially tapered field solenoid that was constructed for an advanced prototype of a mm wave source invented at ETDL. This device requires a solenoidal field that tapers from approximately 1000 oersteds, where the electron beam enters the interaction chamber, to 500 oersteds at the collector end of the chamber. Also, shown is a comparison of the measured field dependence on axial distance with the calculated dependence. While the compromises and approximations made in construction (e.g., substitution of steps for a continuous taper in the outer magnet) result in some deviation from the desired field, agreement is sufficiently good for operation of the tube. The permanent magnet weighs forty pounds and is to replace an electric solenoid which, together with its power supply, weighs over six hundred pounds. the solenoid is also tied to a current source and consumes considerable electrical energy. Furthermore, its stray magnetic field is strong at a considerable distance from the structure; while for the permanent magnet source, the field is largely confined and therefore affords closer packing of field sensitive instrumentation in its vicinity. The permanent magnet structure clearly affords an enormous enhancement in mobility, efficiency, convenience and tractability so that devices formerly confined to fixed stations, ships and large surface vehicles can now be employed in airborne, ballistic and highly mobile surface devices. Permanent magnet solenoids in various stages of design at ETDL include:

a) A field source for an advanced, compact, high-power gyro-amplifier for near-mm waves, designed by ETDL at the request of a major manufacturer.

b) A source for a bi-chambered gyrotron.[7] This source produces a field of 2000 oersteds in the larger chamber and 500 oersteds in the smaller. Through techniques developed at ETDL, this can be accomplished in an abrupt step in field at the juncture of the two chambers.

c) A field source for an extended interaction amplifier to replace a cumbersome, power-consuming electromagnet was designed at the request of another manufacturer.

d) A permanent magnet source to lighten cross-field and extended interaction amplifiers.

e) A klystron magnet for a microwave source for a free electron laser amplifier.

f) A field source for a satellite-borne X-Ray/UV Telescope for NASA.

None of these structures would be viable with conventional magnet materials. This is illustrated in Figure 7. Permanent magnet solenoids all contain magnets operating at point B=0, H=${}_BH_C$ where ${}_BH_C$ is the coercivity. Since the mass and volume of a permanent magnet solenoid are roughly inversely proportional to the square of its coercivity, it is clear that use of Alnico or similar materials would result in a prohibitively bulky structure, two orders of magnitude heavier and larger than Sm-Co magnets producing the same field.

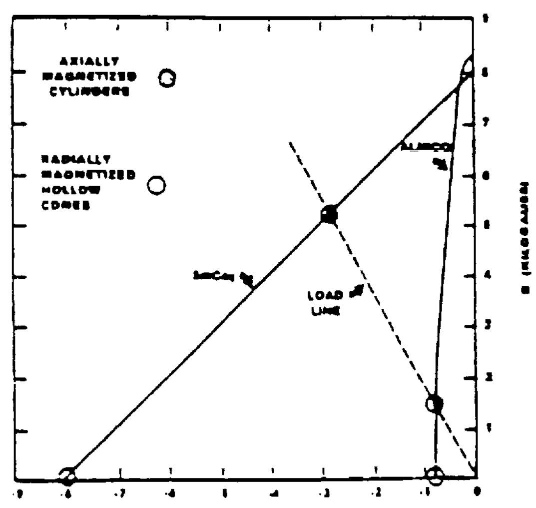


Figure 7. Operating points for permanent magnet solenoid.

Transverse Field Sources

Cylindrical structures that produce fields in interior cavities transverse to their axes are illustrated in Figures 8-13. As in the permanent magnet solenoids, the flux in these structures is confined to their interiors.[3,9-12] The structures in Figures 9-13 also confine flux to their interiors, but employ no iron pole pieces which help "smooth out" small field distortions engendered by structural defects incurred in the course of manufacture and assembly. They can, however, provide much larger fields than structures with iron poles and are limited in field only by the practical considerations of allowable bulk and weight. Figure 13 shows an adjustable version of Figure 12. Field variation is effected by dividing the basic cylinder into two nested rings, each of which contributes the same field to the interior. When the rings are rotated by the same angle in opposite directions, their vector sum will be in the same direction as the original field but smaller. In this way, any field in the range $\pm B_{max}$ can be supplied to the interior without the use of electric currents. All the configurations of Figures 8-13 have potential use in magnetic resonance imaging, or in any application in which uniform transverse fields of thousands of gauss are needed. They are of special value in medical diagnostics because of their small bulk relative to superconducting magnets and their freedom from power sources, cryogens and energy expenditure.

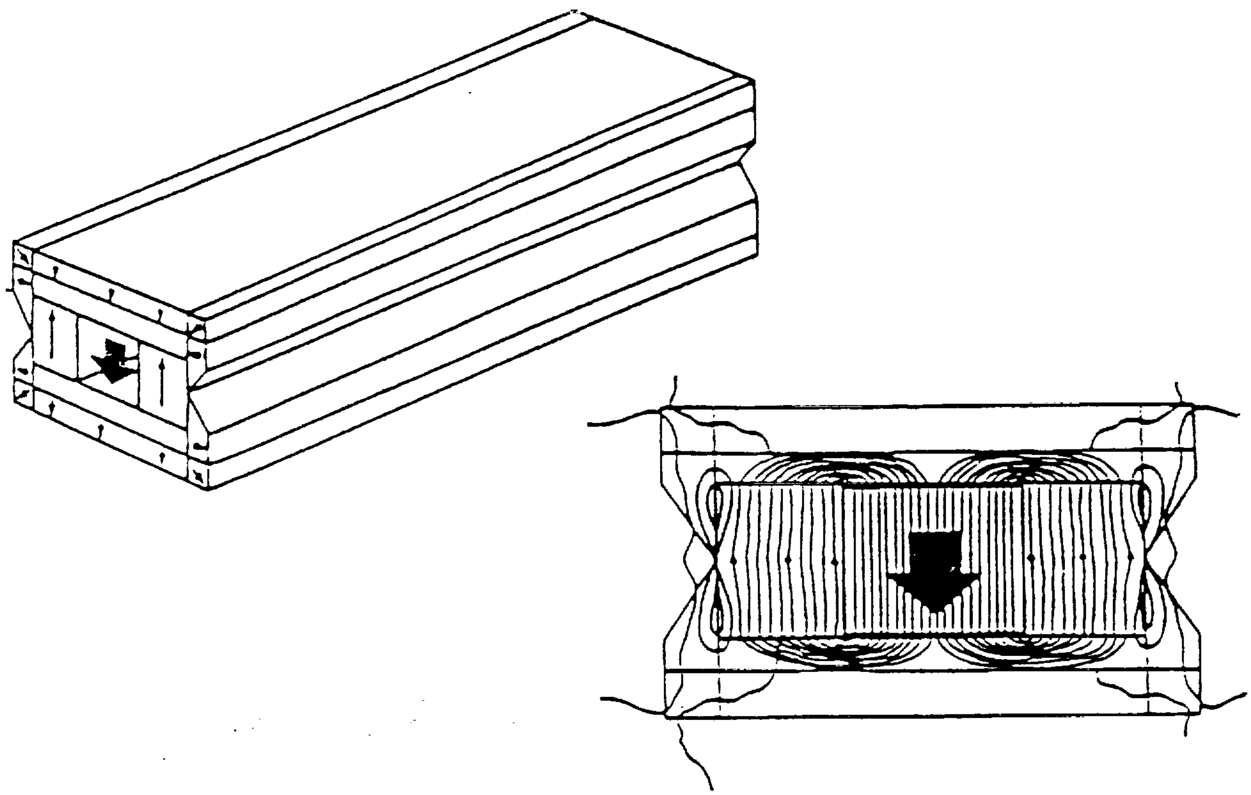


Figure 8. Structure with uniform transverse field in a rectangular working space. The large arrow shows the working field direction and the small arrows show magnet orientations. Possible uses are in NMR imagers and bases for twister structures. Note the field uniformity in the computer flux plot of the cross section.

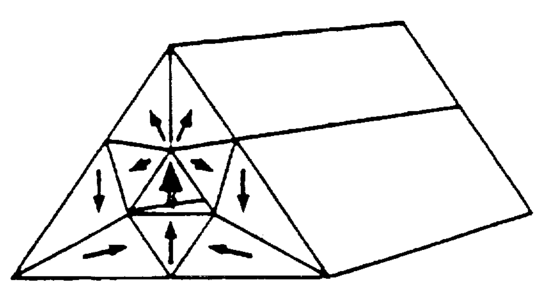


Figure 9. This structure produces a field of one-half the remanance in a triangular cavity. By successive nesting of many of these structures, arbitrarily high fields are attainable.

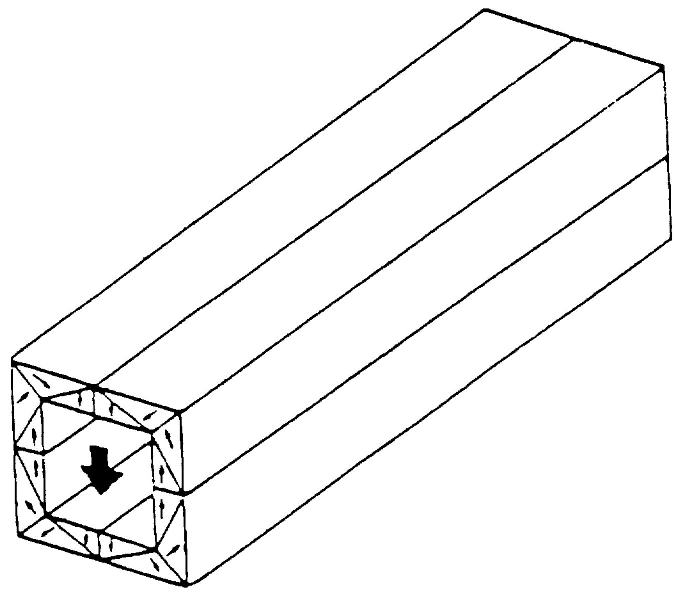


Figure 10. Square permanent magnet structure. As with the rest of the structures on this page, flux is confined to the interior and field augmentation can be attained by sequential nesting.

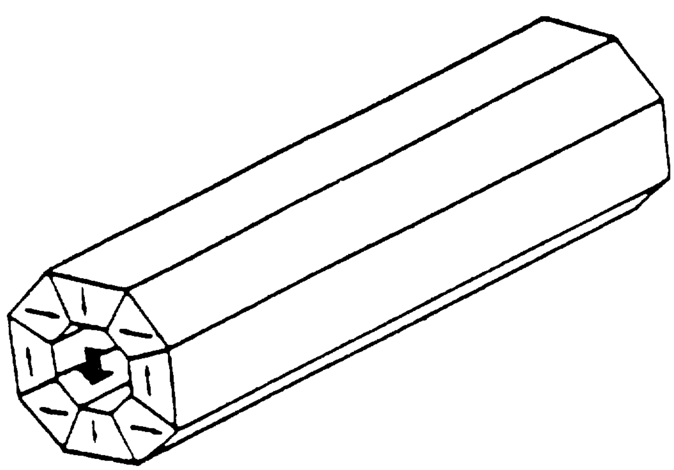


Figure 11. Octagonal dipole structure. Fields attained in these structures are 90 percent of those of the ideal circular structure of Figure 12.

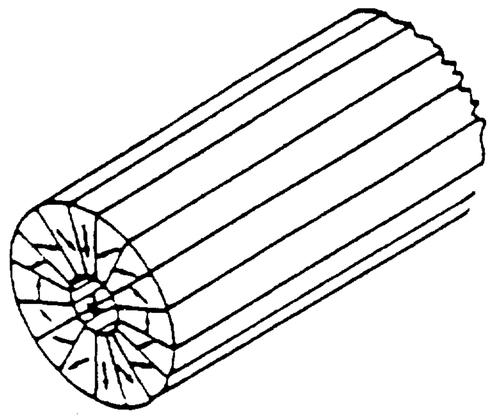


Figure 12. The magnetic orientation (small arrows) varies continuously as 2Θ. A dipolar field (large arrow) is thereby produced in the interior cylindrical cavity of magnitude $H = B_r \ln(P_o/P_i)$ P_o and P_i are the outer and inner radius of the annular magnet.

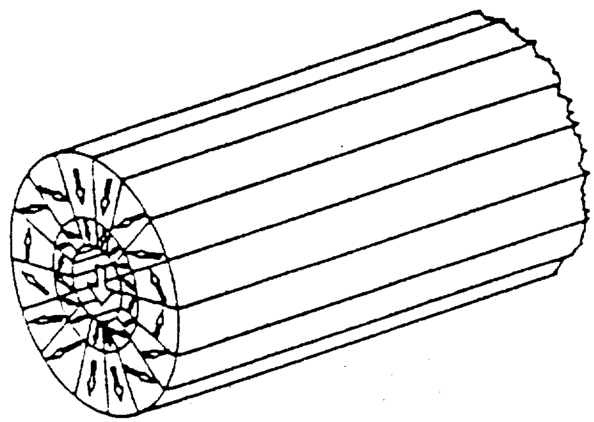


Figure 13. The inner and outer cylindrical shells of the above structure produce the same field H_o in the cavity. Therefore, rotation of the rings with respect to each other yields any field between $\pm 2H_o$ in the cavity.

At present, these drawbacks limit MRI's to large, wealthy institutions where an entire suite can be dedicated to their use. The structures that employ permanent magnets can be made much more cheaply and afford a degree of portability not attainable with superconducting magnets. Therefore, military field use down to divisional or even brigade level is not inconceivable. Moreover, the permanent magnet structures lend themselves readily to miniaturization so that much smaller and more mobile systems could be made for the examination of human extremities and heads. The same miniature systems could also serve as pedagogical devices for the quick training of large numbers of MRI technicians. Applications to anti-terrorist and anti-drug activities also seem feasible; most notably for baggage inspection for contraband at airports and harbors. Several small MRI magnets have already been built and one of half-body size is in the process of assembly and adjustment.

Periodic Structures

A third set of magnetic configurations is formed from cross sectional slices of the second set arranged in periodic arrays as in free electron lasers (FEL) such as wigglers and twisters (Figures 13, 14 and 15). Such FEL arrays produce higher fields than conventional configurations of similar period, beam diameter and structural mass. They can also be corrected for small field distortions arising from dipoles that must be placed in the inner corners of the polygonal structure to effect the desired compensation. The procedure is particularly valuable in FEL's which are notoriously difficult to adjust. Adjustment is further facilitated by use of the bi-ringed nested structure mentioned above.

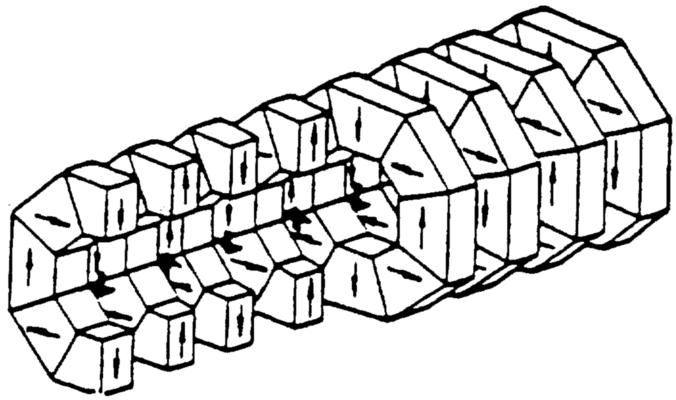


Figure 14. Wiggler structure composed of sectional slices of configuration 11.

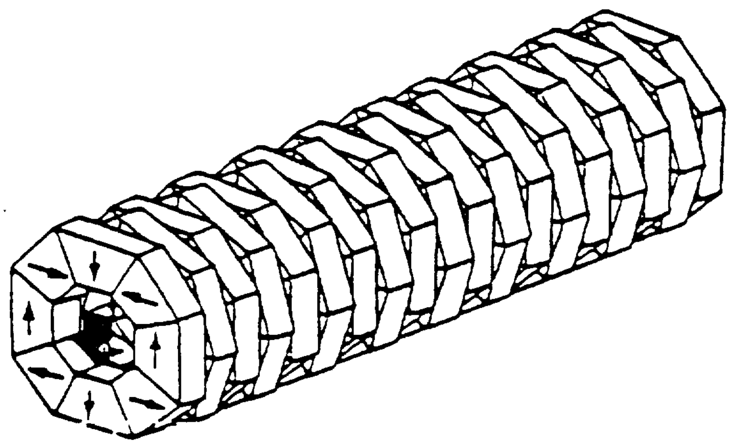


Figure 15. Twister structure composed of sectional slices from the configuration of Figure 11.

In FEL's, an electron beam is sent through a transverse magnetic field that changes directions in a periodic fashion. If the field alternation is roughly sinusoidal as in Figure 14, the electron beam is accelerated from side to side in a direction normal to both its translational motion and the applied field and hence, the device is called a wiggler. The acceleration caused the electrons to radiate energy at the frequency determined by electron velocity and wiggler period. If the proper relationship between velocity, field strength and period exists, the radiation from all parts of the wiggler reinforces and a type of laser action results. When such a relationship exists, the wiggler is called an undulator. If the field remains constant in magnitude but rotates continuously along the structural axis, the electrons are made to follow helical rather than sinusoidal paths and the emitted radiation exhibits circular rather than plane polarization. Such a structure is called a twister or helical free electron laser. Circularly polarized radiation gives radars certain enhanced discriminatory properties compared to plane polarized radiation and is therefore preferable for certain military applications. At the request of the Naval Research Laboratory, a simple permanent magnet twister structure was designed at ETDL that replaces a ponderous electromagnetic field source and its power supply with the usual advantages as is illustrated in Figure 16. A prototype was constructed and found to produce the calculated field of 1200 oersteds as compared with the 500 oersteds delivered by the present coil with the dissipation of 200 amps of electricity.

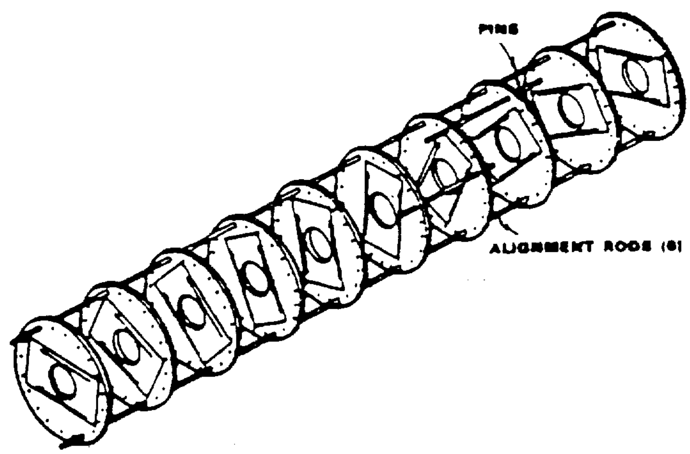


Figure 16. Exploded view of twister structure.

Traveling wave tubes are an important source of mm and microwave radiation for radars and radios. In such tubes, the electron beam is guided by a periodic axial field. Such fields are usually provided by an array of torroidal magnets whose magnetizations are axially oriented in alternate directions and which are interspersed with iron pole pieces that lead magnetic flux into the working space (see Figure 17a). Since such structures are often used in airborne and ballistic devices, where weight and bulk are critical, it is of paramount importance to keep them as light and compact as possible. This is especially true for miniature remotely piloted airborne vehicles where every pound of weight saved can translate into a considerable increase in range or effective payload.

Structures 17b, 17c and 17d were designed and constructed by ETDL to fulfill this need for lighter TWT structures.[14,15] The configuration of 17b produces the same field as that of 17a with one half as much material while those of 17c and 17d result in from one to two order-of-magnitude mass reductions. The latter two arrays, however, are more difficult to manufacture and adjust and so are not so desirable as 17b except in applications with the most stringent of mass limitations. Structure 17c and 17d are very similar with regard to field to mass ratio and differ mainly in that the focusing ability of 17d is enhanced by a stronger field gradient that exists between its axis and the inner walls of the tube. Both tubes exhibit much better flux confinement than does the conventional structure 17a.

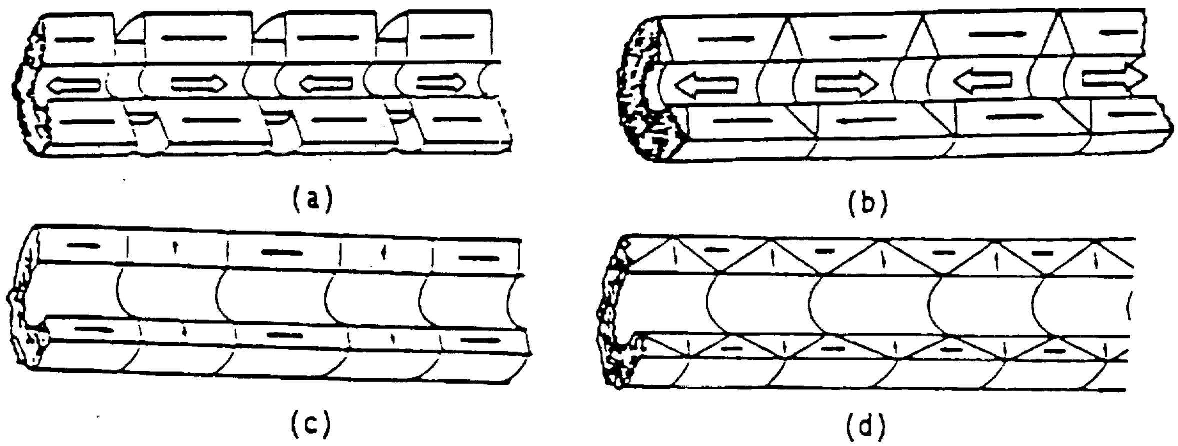


Figure 17. A) Conventional periodic permanent magnet stack for traveling wave tubes. Iron rings are sandwiched between the axial magnetic guide flux into the bore. B) A stack with triangular pole pieces. C) An all permanent-magnet stack consisting of both radially and axially magnetized rings producing the same field as (A) with 1/20th the mass and bulk for a field amplitude of 5.0 kOe. D) A stack with similar performance to that of (C) but with a larger field gradient in the bore.

Very High Field Structures

The fourth class of novel structures is generated by rotations of laminar sections of the second class about their polar axes. Figures 9-12 depict representative examples. Such structures produce, in their interior cavities, the very highest fields presently obtainable with permanent magnet structures. For example, in a spherical structure with an inner diameter of one inch, a field of two tesla is obtainable with an outer diameter of 4.5 inches if a remanence of 10 kG is used. With a Nd-Fe-B magnet of 12 kG remenance, the same field is obtainable with an outer diameter of only 3.5 inches.

If a structure of this type is cut in half at the equator and placed on a planar passive ferromagnet such as iron or permendur, the anti-mirror image formed in the ferromagnetic plane produces the same field as the missing half of the original sphere. Such a system is easier to manufacture as it requires only half as many pieces. It also provides more convenient access to the interior as no holes need be bored through the expensive permanent material and different access configurations are easily obtainable in the same hemisphere with different iron plates thereby adding greatly to versatility of use (see Figure 18).

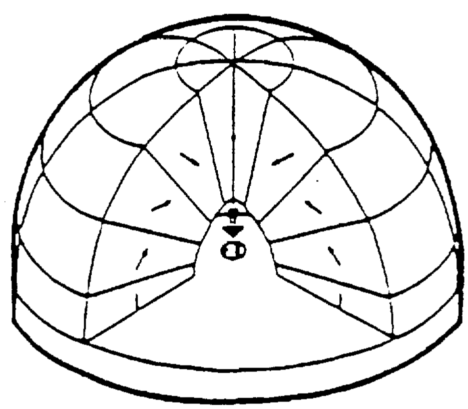


Figure 18. "Magic Igloo," hemispherical magnet set on iron plate.

These structures are useful where very high fields are required such as in some Faraday rotators and short travel beam tubes. Such structures placed in tandem may also be useful as wiggler elements when placed with poles normal to the electron-beam axis or as traveling wave tube elements when arranged in tandem, parallel to the beam axis.

REFERENCES

1. H. A. Leupold, E. Potenziani II, and J. P. Clarke, Proc. Ninth International Workshop on Rare-Earth Magnets, Bad Soden, West Germany, p. 109, (1987). Available from Dr. K. Strnat, E. E. Dept. of University of Dayton, Dayton, OH.

2. H. A. Leupold, E. Potenziani II, J. P. Clarke, and D. Basarab, Materials Research Society, Symp. Proc. 96, p. 279 (1987).

3. J. J. Clarke and H. A. Leupold, IEEE-Mag, Trans. of the Magnetics Society, Mag-22, No. 5, p. 1066.

4. H. A. Leupold and E. Potenziani II, IEEE-Mag, Trans. of the Magnetics Society, Mag-22, No. 5, p. 1078 (1986).

5. J. J. Clarke, E. Potenziani II, and H. A. Leupold, Jour. of Applied Physics 61, (8), p. 3468 (1987).

6. H. A. Leupold, "Powder Metallurgy in Aerospace and Defense Technologies," p. 167 (1990).

7. H. A. Leupold, E. Potenziani II, and A. Tilak, Jour. of Applied Physics 67, p. 4650 (1990).

8. J. Lowrance, C. R. Joseph, H. A. Leupold, and E. Potenziani II, to be published in "Advances in Electronics and Electron Physics" (1991).

9. M. G. Abele and H. A. Leupold, Jour. of Applied Physics 64, p. 5988 (1988).

10. H. A. Leupold, E. Potenziani II, and M. G. Abele, Jour. of Applied Physics 64, No. 10, p. 5994 (1988).

11. M. G. Abele, 10th International Workshop on Rare-Earth Magnets and Their Applications, Kyoto, Japan, 16-19 May 1989 (Proceedings Book: The Society of Non-Traditional Technology, 1-2-8 Toranomou Minato-ku, Tokyo, 105 (Japan).

12. K. Halbach, Proc. of the Eighth International Workshop on Rare-Earth Magnets and Their Applications, Dayton, OH, p. 123 (1985).

13. A. B. C. Marcos, H. A. Leupold, and E. Potenziani II, IEEE-Mag. Trans. of the Magnetics Society, Mag-22, No. 5, p. 1066 (1988).

14. H. A. Leupold and E. Potenziani II, Jour. of Applied Physics 67, p. 4650 (1990).

15. H. A. Leupold, E. Potenziani II, and A. Tilak, to be published in Proc. of IEDM (1991).

ADVANCED MANUFACTURING

(Session B1/Room A1)

Wednesday December 4, 1991

- Concentrating Solar Systems: Manufacturing With The Sum
- Ultra-Precision Processes for Optics Manufacturing
- Integrated Automation for Manufacturing of Electronic Assemblies
- Air Force Manufacturing Technology (MANTECH) Technology Transfer

N92-22448

MANUFACTURING WITH THE SUN

Lawrence M. Murphy
Steven G. Hauser
and
Richard J. Clyne

National Renewable Energy Laboratory
(formerly the Solar Energy Research Institute)
1617 Cole Blvd.
Golden, CO 80401

ABSTRACT

Concentrated solar radiation is now a viable alternative energy source for many advanced manufacturing processes. Researchers at the National Renewable Energy Laboratory (NREL) have demonstrated the feasibility of processes such as solar-induced surface transformation of materials (SISTM), solar-based manufacturing, and solar-pumped lasers. Researchers are also using sunlight to decontaminate water and soils polluted with organic compounds; these techniques could provide manufactures with innovative alternatives to traditional methods of waste management. The solar technology that is now being integrated into today's manufacturing processes offers even greater potential for tomorrow, especially as applied to the radiation-abundant environment available in space and on the lunar surface.

INTRODUCTION

Researchers at the Department of Energy's (DOE's) National Renewable Energy Laboratory (NREL) are developing advanced manufacturing processes that are powered by the sun. These processes are offering industry viable alternatives for accomplishing many of today's manufacturing tasks, and they are likely to be a key ingredient in solving energy problems in tomorrow's advanced manufacturing environment. The advanced solar manufacturing research is managed through NREL's Mechanical and Industrial Technology Division, which is dedicated to bringing American industry practical renewable alternatives that reduce or replace fossil-fuel-based sources of energy.

Currently, the most promising results related to materials manufacturing have come from using concentrated solar radiation to modify the surface properties of materials used in advanced technology applications. Applications such as transformation hardening of steel, cladding, self-propagating high-temperature synthesis, thin-film deposition, and electronic materials processing have been successfully demonstrated and analyzed at NREL.

The advanced manufacturing research is centered around NREL's high-flux solar furnace, which can achieve solar intensities of up to 50,000 times normal ambient conditions. While using directed energy beams for advanced manufacturing has been understood and integrated in certain applications for more than a decade, researchers are now demonstrating that solar-energy-based processes are economically competitive with traditional methods of materials processing [1].

These processes are also environmentally benign, as all the energy driving them comes directly from the sun. High levels of energy efficiency are achieved because the energy conversion steps inherent in traditional processing—fossil fuel to heat, heat to electricity, electricity back to heat—are eliminated. Reflected, concentrated solar flux is applied directly to the materials being processed.

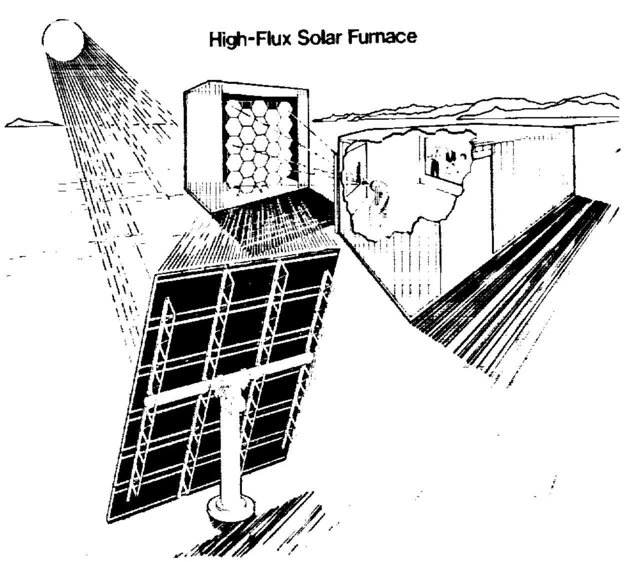
NREL researchers are also investigating solar-based methods to reclaim process water contaminated with organic compounds. This solar detoxification process uses sunlight in combination with a photocatalyst to destroy aqueous volatile organic contaminants. In the near future, as early as the middle of this decade, this process could be a viable waste management technique to cleanse water used in manufacturing processes and to reclaim polluted water.

NREL's HIGH-FLUX SOLAR FURNACE

New techniques for manufacturing advanced materials are being explored through the use of NREL's solar furnace. The solar furnace began operating in early 1990 and allows researchers to study the properties and applications of very high solar flux. The facility currently has the ability to produce solar flux densities of up to 50,000 suns (5000 W/cm^2). This upper limit rose dramatically recently with a new first-of-its-kind concentrator design. New research directions and applications for this promising technology are continually being discovered.

Figure 1 shows the physical configuration of the solar furnace. The heliostat tracks the sun and reflects incoming solar energy onto the stationary primary concentrator, which consists of 23 individual curved facets (see Figure 2). These facets collectively focus the solar flux at a point in the test facility, which is located just off the primary axis. The long focal length of the primary concentrator produces a 10-centimeter-diameter concentrated beam of approximately 2,500 suns at the center of the target area. When a secondary concentrator is placed at the beam's focus, the peak flux can be increased 10 to 20 times, the upper limit applying to dielectric-filled secondaries.

Three performance characteristics put the solar furnace at the cutting edge of solar industrial process research. The first is the system's ability to produce very high temperatures directly from the sun. (For example, we have melted through a 2-inch alumina fire brick, which has a melting point of 1800°C, in less than 1 minute.) The second is the extremely high rate of heating made possible by very high solar flux. In certain applications this rate exceeds 1 million degrees Celsius per second. These heating rates produce thermal conditions attainable only through radiant heating processes. The third characteristic is the furnace's ability to deliver the entire solar spectrum (from 300 to 2500 nanometers); this allows researchers to study applications requiring either broad spectrum radiation or a particular frequency, ranging from the infrared to the near-ultraviolet.



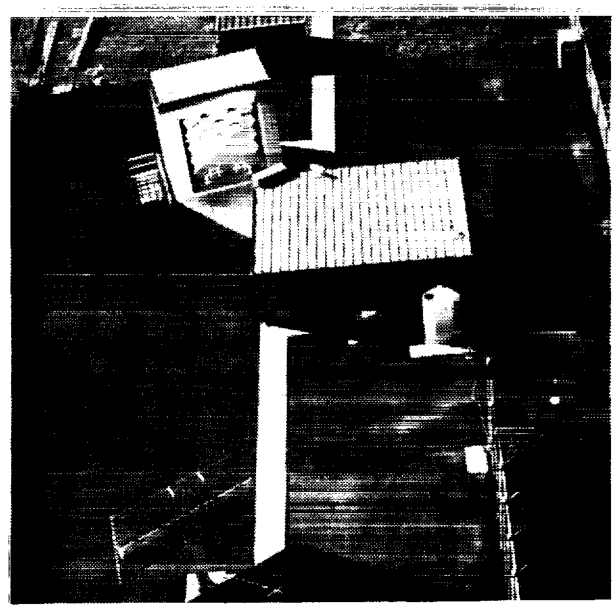


Figure 1. NREL's high-flux solar furnace.
Left: Schematic view of system operation.
Right: The actual facility located on South Table Mountain in Golden, Colorado.

MATERIALS PROCESSING APPLICATIONS

Materials processing is a major area of research at NREL's high-flux furnace facility. Specifically, research pertaining to the solar-induced surface transformation of materials (SISTM), which is ideally suited to the high-flux environment created by the solar furnace, is at the center of these studies. SISTM research uses the thermal energy (as opposed to the photochemical energy) produced from solar radiation. The key to SISTM is the rapid, controlled heating that alters the surface of a workpiece without affecting its base properties. The processes we are developing can be carried out without consuming fossil resources and without producing a host of environmental liabilities.

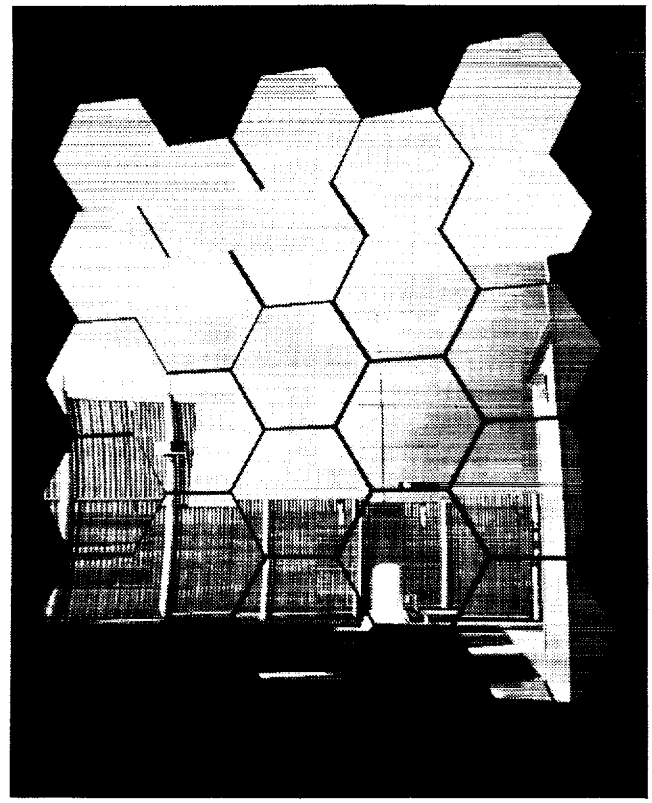


Figure 2. The primary concentrator of NREL's solar furnace. The 23 enhanced-aluminum front-surface mirrors focus the reflected solar flux to a 10-cm-diameter beam inside the test building (not shown). The reflection of the heliostat, which tracks the sun and directs its energy to the concentrator, is shown.

NREL researchers are refining ways to use solar energy to produce surface modifications that are critical to a number of materials technologies including hardening, cladding, chemical vapor deposition, and the manufacture of electronic components and circuitry. These are currently being fabricated using electron beams, arc lamps, lasers, and induction heating. Preliminary economic analyses indicate that concentrated solar flux, when used in large-scale production applications, could produce these materials at one-half to one-quarter of the cost of production associated with the conventional radiant methods in use today.

The transformation hardening of steel is typically accomplished using lasers and is being applied most noticeably in the automotive industry for hardening drive-train components. Researchers at NREL have demonstrated that steel can be hardened using solar energy, and have shown that the process is competitive with laser-based techniques [2]. Hardened steel consists of a hard, wear-resistant layer surrounding a softer, steel core. Figure 3 shows a hardened steel sample produced in the high-flux furnace.

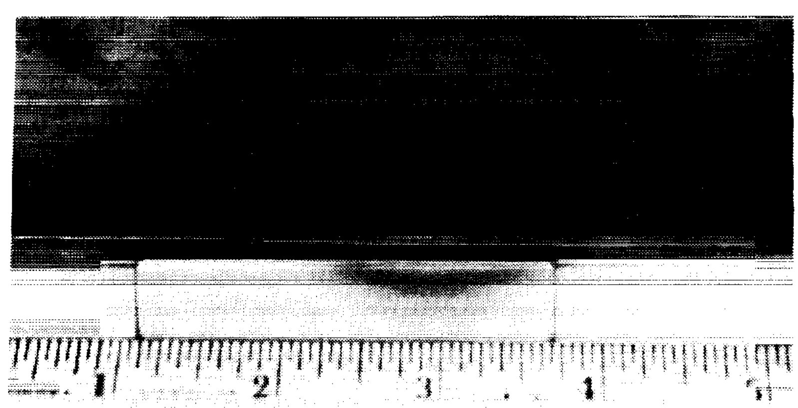


Figure 3. Cross section of solar-hardened steel. The fully hardened region is 1 to 2 mm deep and about 20 mm in diameter.

Using solar radiation to clad preapplied powders to steel substrates has generated a considerable amount of industrial interest. Solar cladding produces excellent

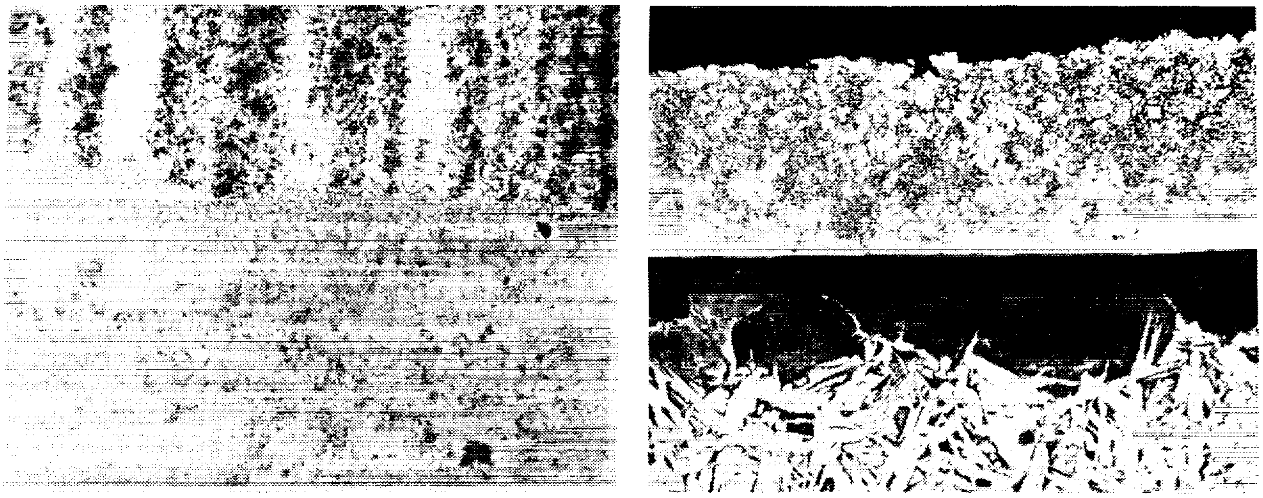


Figure 4. Photomicrographs of alloy coatings formed by solar melting of preapplied powders on steel substrates. Note the high density of the coatings and the good metallurgical bonds between the coatings and their substrates. The images show an area measuring approximately 600 μm by 400 μm.

metallurgical bonds between the melted powder and the substrate. The desirable properties of an expensive material such as a superalloy can be economically obtained by cladding relatively small amounts of this material to a less expensive substrate base such as mild steel. Figure 4 shows micrographs of solar-clad materials.

Concentrated sunlight is being applied to a number of chemical vapor deposition processes. Chemical vapor deposition consists of generating sufficient levels of heat on or near the surface of a given workpiece in the presence of selected gaseous reactants. The heat causes the reactants to form solid products on the surface. Solar furnace technology is well suited to this process because the surface heating can be closely controlled, eliminating the formation of solid product on surfaces other than the specific material of interest. NREL is investigating the chemical vapor deposition process for the production of TiN, TiB_2, SiC, and hard carbon films.

Rapid thermal annealing (RTA), a process that uses radiative heat to accomplish diffusion processes or to remove defects from delicate electronic components such as semiconductors, has traditionally relied on tungsten-halogen lamps as the source of energy. The high level of control required for RTA is easily obtained using concentrated solar radiation; the solar furnace meets the precise requirements of this process. Depending on the materials being annealed, the solar furnace rapidly heats a workpiece to operating temperature and provides the required thermal treatment. RTA studies conducted at NREL's solar furnace include producing high-temperature superconductor films on substrates such as $SrTiO_3$, ZrO_2, and MgO.

SOLAR-PUMPED LASER APPLICATIONS

Recent advances in secondary concentrator systems are causing researchers to reevaluate using concentrated solar flux to pump lasers. The concept of solar-pumped lasers has been studied for more than two decades, but low beam concentrations limited the conversion efficiencies to approximately 1%. With the upper bounds of attainable concentration now approaching 50,000 suns, conversion efficiencies are approaching 5%, which offers the potential for both high power and high efficiency for several types of lasers [3].

A solar-pumped laser would have all the performance characteristics of traditionally powered devices, and a demonstration of a 5%-efficient solar-pumped laser should occur in mid-1992. Researchers are developing innovative applications for solar-pumped lasers, such as the destruction of toxic compounds (polychlorinated biphenyls, PCBs) and the manufacture of inorganic high-value substances (ceramic carbides and borides) [3].

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

DESTROYING CHEMICAL WASTES

NREL researchers are studying two distinct solar-based methods to destroy hazardous environmental waste. The first process—the solar detoxification of water—uses low solar concentrations to breakdown organic compounds in contaminated process water and groundwater. The second method—known as gas-phase detoxification—uses the concentrated solar flux produced in the solar furnace to destroy hazardous compounds that have been desorbed from solids and exist in the gas phase.

The first process destroys unwanted organic compounds found in waste water or polluted groundwater. The process works as follows: polluted water is brought into contact with a semiconducting photocatalyst, such as titanium dioxide, which in the presence of sunlight creates highly reactive hydroxyl radicals. These radicals react with the organic compounds and convert them into water, carbon dioxide, and easily neutralized mineral acids such as hydrochloric acid. Because the process requires low solar flux concentrations, the reaction can be initiated using either flat-plate solar panels (one-sun concentrations) or a transparent tube mounted at the focus of a parabolic reflecting trough (up to 30-sun concentrations). This latter design is shown in Figure 5. This process is currently being field tested at a Superfund site and could be available to industry as a waste management tool soon.

Figure 5. Experimental solar decontamination trough.

Gas-phase detoxification is a hazardous waste destruction method being applied to contaminated solids, such as soils. This process, which is being studied at NREL's high-flux solar furnace, can use energy from throughout the solar spectrum. Typically, a concentration of greater than 300 suns is required to drive the reaction. During the process the contaminants are desorbed from the solid using either vacuum extraction or heat (heat produced by the infrared portion of the solar spectrum can be used for this purpose). After being desorbed, the contaminants are introduced into a reactor (shown in Figure 6) that is mounted at the focus of a solar concentrator system. The near-ultraviolet portion (high-energy photons) of the flux then converts the contaminants to end products such as CO_2, H_2O, and HCl. Catalysts are being studied that can greatly improve the effectiveness of the process.

Both of these methods are remarkably efficient in destroying hazardous organic compounds (exceeding the EPA's 99.9999% requirement), and neither produce the detrimental by-products associated with many of the conventional waste remediation techniques [4]. By the latter part of this decade these methods could become standard tools for decontaminating process waste water, soils, and other contaminated media.

SOLAR MANUFACTURING PROCESSES IN SPACE

The concentrated solar radiation applications that are now being proven and integrated into manufacturing processes here on earth hold even greater potential for use in space. As we strive for a permanent presence in space, the practical issues of energy logistics and efficiency will be crucial to the success of the endeavor. Solar technology is remarkably well suited to supply a large portion of the energy needed in a working space environment.

Concentrated solar radiation technology would produce greater efficiencies in space than attainable in the terrestrial environment. Available solar radiation increases by 70% outside the earth's atmosphere [5], and the ultraviolet portion of the spectrum expands down to 200 nm, providing more energy for photolytic interactions. Also, the direct use of solar radiation would employ a much smaller solar-collecting system because of the high efficiencies inherent in the technology. For example, to attain the same levels of usable power as a direct solar beam system, a photovoltaic-driven arc lamp, which experiences high energy losses converting solar radiation to electricity, would require 10 times the amount of collector surface area; a photovoltaic CO_2 laser system operating at the same level would require a collector 25 times larger than that used for a direct solar beam system. Additionally, the space environment would introduce complications in dissipating the excess heat produced by these conventional technologies [6].

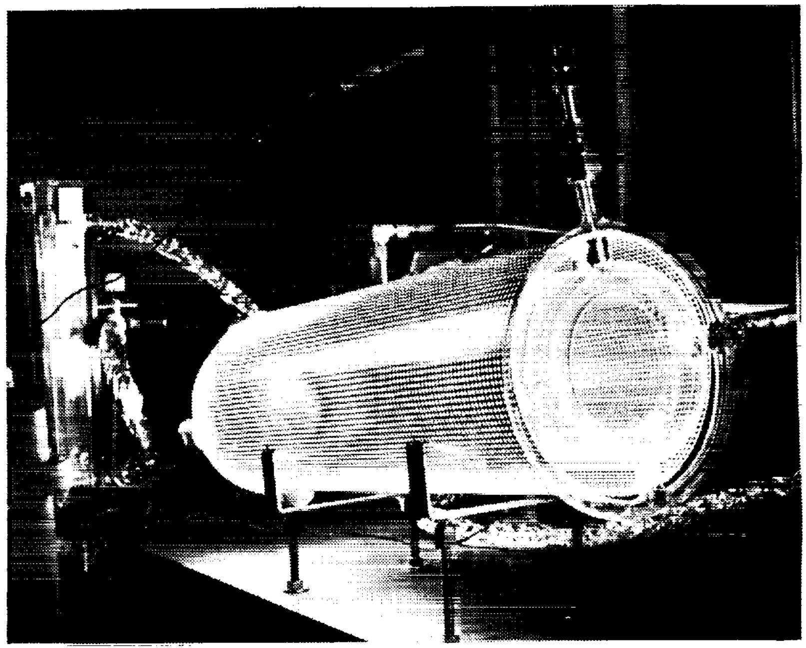


Figure 6. Solar gas-phase detoxification reactor.

In addition to the surface modification applications discussed previously in the Materials Processing section of this paper, a number of other applications are readily identified for using concentrated solar radiation in a space environment. High solar flux could be applied to materials joining, welding, fabricating, repairing, and surface cleaning operations. The technology could also be integrated in lunar mining operations to provide bulk heating for extracting certain products from mined bulk materials. An artist's conception of a lunar-based solar furnace is shown in Figure 7.

Water would be available in limited quantities in any space environment, and solar radiation could be applied to extract water from human waste and reclaim water used in operating processes.

RESEARCH COLLABORATION AND TECHNOLOGY TRANSFER

The long-term success of NREL's research and development projects is tied to working closely with industry and other outside research organizations.

NREL works with industrial partners through a variety of arrangements, including cost-shared demonstrations, joint research, and Cooperative Research and Development Agreements (CRADAs). NREL technology transfer is fostered through written material, presentations, workshops, training programs, and traveling exhibits. Our participation in industry projects, and exchanges between our researchers and their industry counterparts, also play an important role in disseminating technology advances.

Only through close cooperation and information exchange among researchers, manufacturers, and operators will solar-based manufacturing processes continue to make inroads into U.S. industry.

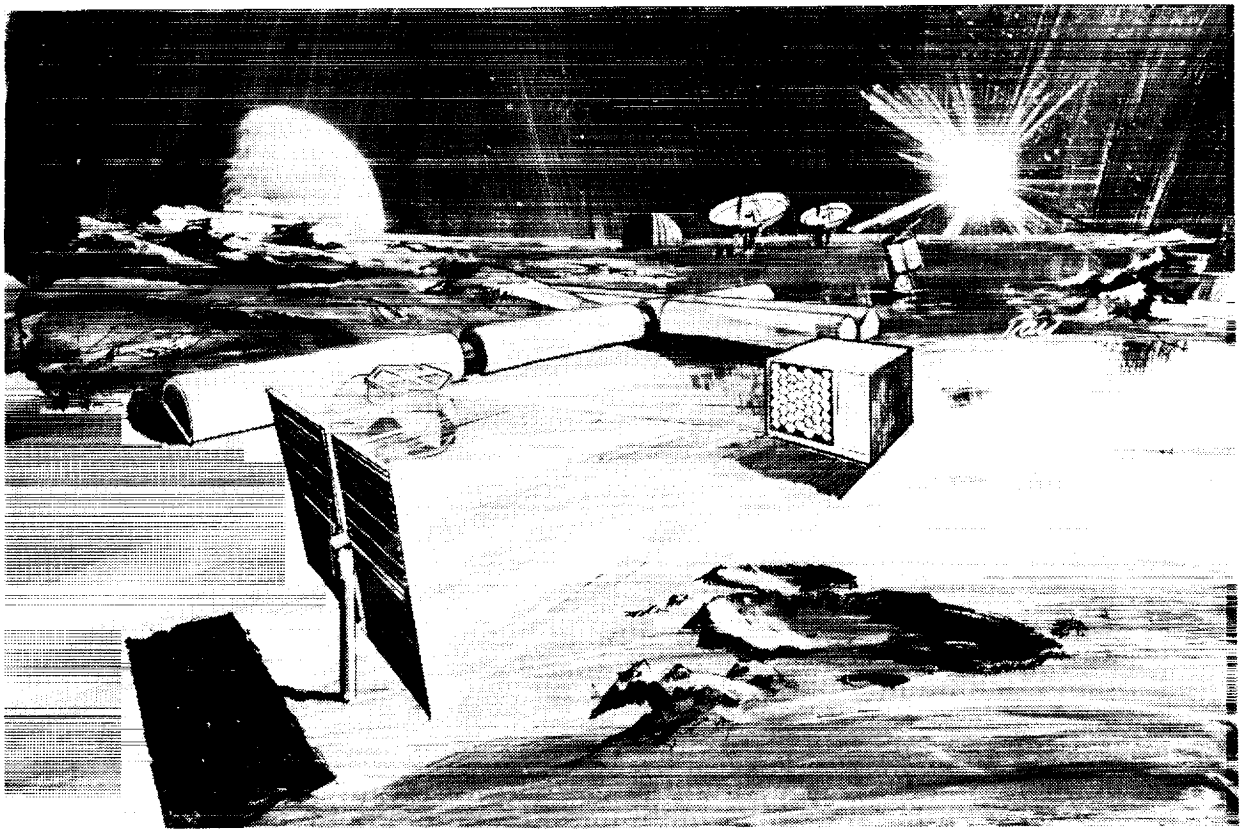


Figure 7. Artist's conception of lunar-based solar furnace.

CONCLUSION

The solar-based manufacturing processes developed at NREL have advanced to the point where they can be considered competitive energy alternatives for many conventional manufacturing applications. One of the most attractive features of solar manufacturing systems is their immunity to fluctuations in the price of fuel; operation and maintenance costs remain extremely stable over the life of a facility.

As this technology gains a foothold in the marketplace, other solar processes currently in the developmental stage will mature. Certainly these advances will play a crucial role in the next generation of advanced manufacturing technology, both on earth and in space.

REFERENCES

1. Pitts, J.R., C.L. Fields, J.T. Stanley, and B.L. Pelton, "Materials Processing Using Highly Concentrated Solar Radiation," Proceedings of the 25th Intersociety Energy Conversion Engineering Conference, Volume 6, American Institute of Chemical Engineers, 1990.

2. Stanley, J.T., C.L. Fields, J.R. Pitts, "Surface Treating with Sunbeams," Advanced Materials and Processes, Vol. 138, no. 6, December, 1990.

3. Rosen, J., "Solar Furnaces: Concentrating 100,000 Suns," Mechanical Engineering, February, 1990.

4. Gupta, B.P., J.V. Anderson, Solar Detoxification of Hazardous Waste: An Overview of the U.S. Department of Energy Program, Solar Energy Research Institute, SERI/TP-253-3959, August 1990.

5. Riorden, C.J., D.R. Meyers, R.L. Hulstrom, Spectral Solar Radiation Data Base Documentation, Volume II, Solar Energy Research Institute, SERI/TP-215-3513B, July, 1989. This figure assumes 1367 W/m^2 mean direct normal radiation at earth orbit and air mass 0, and 800 W/m^2 direct normal availability at air mass 1.5 for the southwestern United States.

6. Pitts, J.R., T. Wendelin, J.T. Stanely, "Applications of Solar Beams for Materials Science and Processing in Space," Proceedings of the 25th Intersociety Energy Conversion Engineering Conference, Volume 1, American Institute of Chemical Engineers, 1990.

N92-22449

ULTRA-PRECISION PROCESSES FOR OPTICS MANUFACTURING

William R. Martin
Oak Ridge National Laboratory
Oak Ridge, TN 37831

ABSTRACT

The Optics MODIL (Manufacturing Operations Development and Integration Laboratory) is developing advanced manufacturing technologies for fabrication of ultra-precision optical components, aiming for a ten-fold improvement in precision and a shortening of the scheduled lead time. Current work focuses on diamond single point turning, ductile grinding, ion milling, and in/on process metrology.

INTRODUCTION

Industry in this country does manufacture sophisticated optics. For now the process is laborious, time consuming and the results are difficult to predict. But the opticans are artists and accomplish much. For the next century, there will be better methods.

Technology emerging today may be able to impact the market place in the Year 2001. As the market pulls the technology into industrial applications, that emerging technology will be modified to enable a profitable application. In our country, more and more cooperative and collaborative R&D efforts are in effect between the tripartite, industry, universities, and federal laboratories. We are beginning to better leverage our technological assets to provide the citizens of our country the opportunity to work in high value-added industries that are globally competitive. There are a number of important on-going efforts to improve the producibility of optics. Several Optics Science Centers at our major universities are involved in outstanding efforts. Fortune 500 companies and small entrepreneurial firms are involved in ingenious new approaches. Federal funding is supporting the effort of the Center for Optics Manufacture and the Optics MODIL.

The Optics MODIL, funded by the Strategic Defense Initiative Organization, is integrating the emerging and enabling technology to manufacture ultra precision optics that will be affordable. The size of the ultra precision optics market is currently both small and volatile. There is a lack of current market incentives for industry to accelerate the deployment of ultra precision technology. While one can debate what application will expand this ultra precision market segment, it is important to recognize that the Europeans, especially the Germans, and the Japanese have very strong manufacturing development programs for Nanolevel Finishing Technology. It's probable that those foreign efforts are not just for the pursuit of technology for technologies sake. Commercial application and markets include: analyzers - environmental monitors, air quality; diagnostics - process monitors, scanners; laboratory equipment - spectrometers; IR imagers; satellites - high cost, low quantity; diffractive optical elements; contact lenses, aspheric corrector; precision molds - plastic, optics (fresnel optics), contact lenses; precision machined components for precision machines. While we have mentioned a number of applications, there may be even more important ones that we have missed.

MIRROR MANUFACTURE

The approach of the Optics MODIL is to develop manufacturing processes for ultra precision work that is: deterministic, affordable investment for small and large business, adaptable for in/on process metrology, applicable to a range of materials, and flexible with regard to shape and size.

A desirable feature of this new capability would be the ability to provide custom designs at prices we currently reserve for off-the-shelf procurements. Another view would be to offer the product in low quantities (i.e. 5 to 10) that currently are reserved for total buys of 500 to 1,000. The new business, must be able to take a few designer optical characteristics and build the product. Tooling and fixturing costs must be low and readily available. The customer must be willing to reduce the specification to a minimum without providing the manufacturing business with stacks of documents.

Current finishing processes being developed with industry are single point turning, ductile grinding and ion milling with strong metrology development to provide inspection in- or on-process. These efforts are evolving within the umbrella operation of a Producibility and Validation Test Bed where joint industrial programs can be pursued. Within the Test Bed, Manufacturing Cells contain the individual finishing operations. Materials of interest are metals i.e., Beryllium, SXA composites, Silicon, and Electroless Nickel coated substrates. Ceramic materials of interest are principally versions of Silicon Carbide.

Because these operations are focusing on Nanolevel finishing technology, the manufacturing cells that house these operations are capable of excellent control of the environment. The Cells are shown in Figures 1 and 2. Temperature variations within the inter-compartment can be controlled to within a tenth of a degree fahrenheit. This is only one of the factors important for reproducible operations in the nanolevel regime. The ability to consistently produce products to a few tens or hundreds of nanometers requires consistency of the operations. Equipment required to achieve this type of control is commercially available at about $\$75/ft^2$.

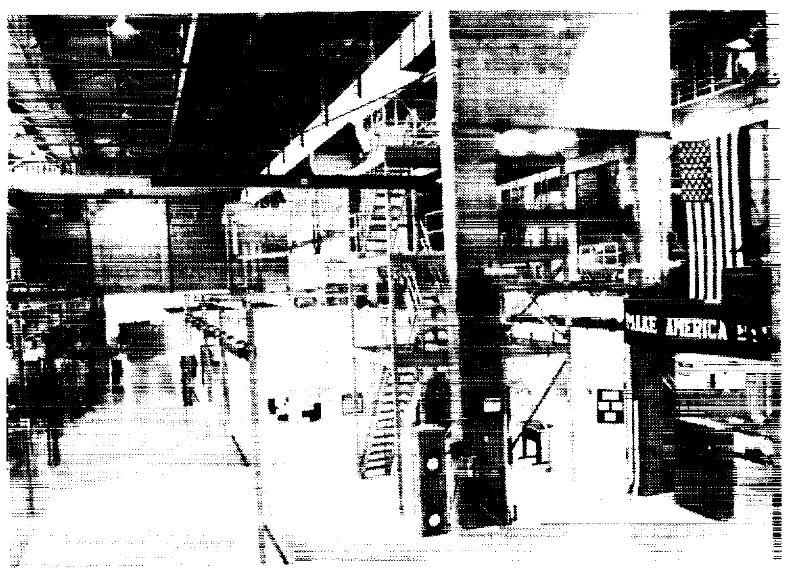


Figure 1: Holonic Manufacturing Cells in Producibility and Validation Test Bed (PVTB)

Figure 2: Interior View of Manufacturing Cells

It is envisioned that these cells would be deployed in businesses such as system houses, special machine shops or perhaps adopted by optical polishing/coating firms as alternative technologies that significantly expand capabilities, lower manufacturing labor, increase yield, and greatly reduce the time necessary to finish the optical surface. A deterministic process that would make optics cheaper, faster, and perhaps better. The overall strategy of developing a manufacturing operation for the finishing process with metal substrates and/or ceramic substrate is shown in Figure 3.

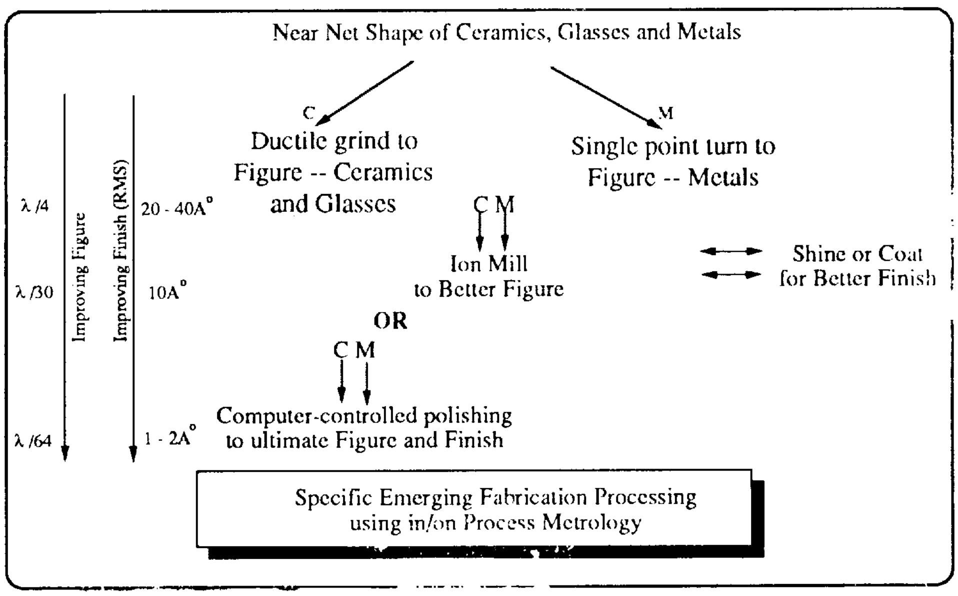


Figure 3: Mirror Fabrication Strategies

Our single point turning efforts utilize a Rank Neumo 600 single point turning machine. This commercially available unit housed in the Productivity and Validation Test Bed (PVTB) is the prototype model developed in this country. Two more units have been sold to U.S. firms and are being delivered by Rank Pneumo during 1991 and 1992. The Japanese are also now buying a unit. A photo of this machine is shown in Figure 4. Mirrors being viewed are shown in Figure 5. We currently have produced Single Point Diamond turned optical surfaces that have figure accuracy of 1/4 to 1/6 wave for small sizes having a Tallystep surface finish about 20 Å RMS. The BRDF light scatter at 10.6 microns was as low as 100 ppm. These as machined surfaces have very low scatter for IR applications. Photos of some as-machined mirrors are shown in Figure 6. To date for small optics, i.e., 155mm diameter F/4 convex hyperbolic mirror, the surface finish is about 25Å RMS for Electroless nickel surface and the figure accuracy has been about one-sixth wave. For larger mirrors, such as a 400 mm diameter, nickel coated SXA foam body, the figure accuracy peak to valley was about one quarter wave (RMS). On the larger parts, centering, fixturing and tooling errors dominate. These data are shown in Table 1. Fixturing schemes are shown in Figure 7. Work by Bob Parks of the University of Arizona has demonstrated that the surface smoothness of these diamond turned parts can be improved by a factor 5 by using a brief flexible lap polish.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

Figure 4: Prototype Rank Pneumo 600 Turning Machine

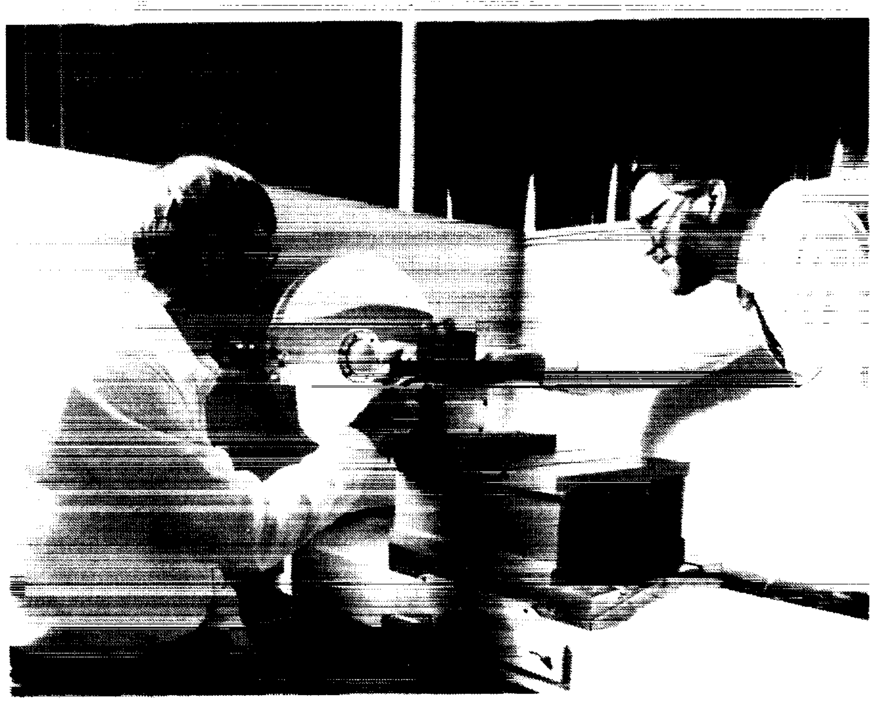


Figure 5: Complex Surfaces of Turned Mirrors are Examined by Staff

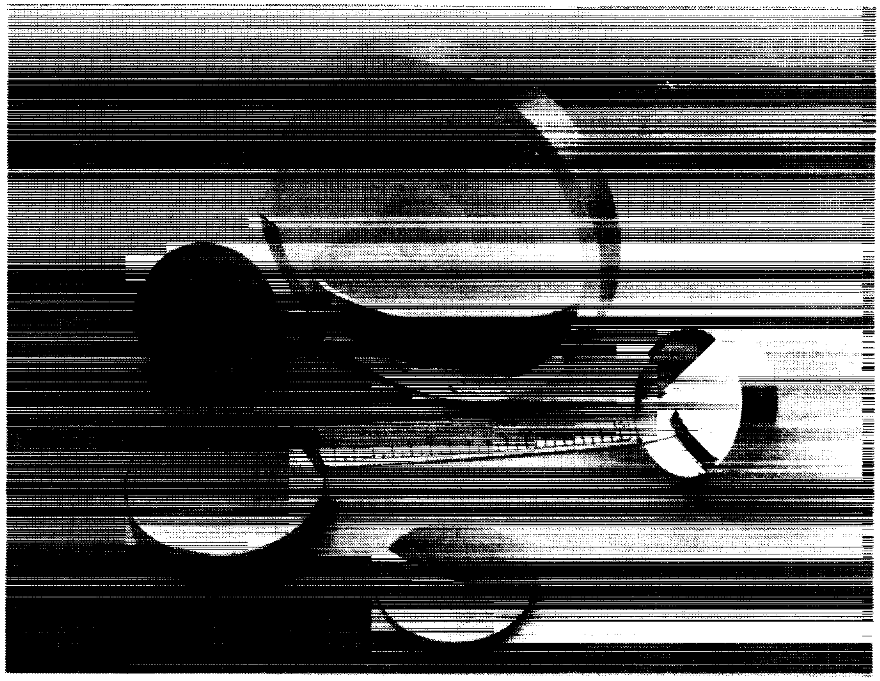


Figure 6: Typical Single Point Diamond Turned Mirrors

Size of Optic mm	Shape of Surface	Peak to Valley Figure Accuracy (visible waves)	Finish			Material
			BRDF			
			10.6	0.633	Å RMS	
40	Flat	0.25	E-4	E-2	---	Aluminum
75	Sphere	0.16	E-5	E-6	20	Copper & Aluminum
155	Hyperbola	1.3	(a)	(a)	25	Electroless Nickel
200	Parabola	0.5	(a)	(a)	---	Aluminum
400	Parabola	1.5 to 0.5	(a)	(a)	---	Aluminum

(a) Currently being deterined

Application in the IR Range

Table 1: Single Point Turning of Metallic Mirrors

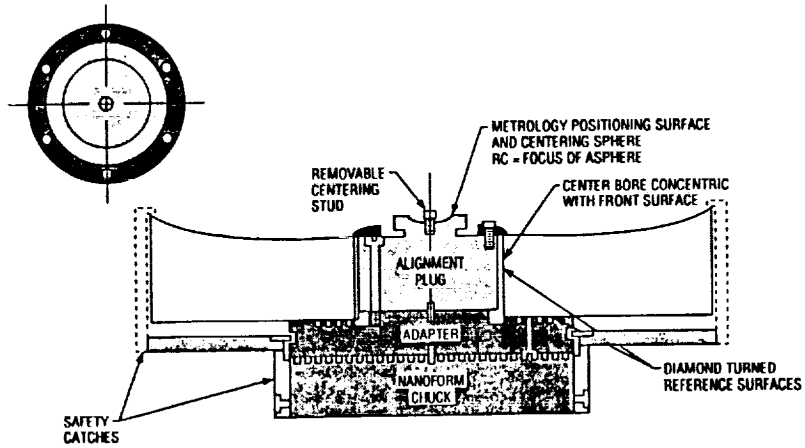


Figure 7: Fixturing Scheme for 400mm Diameter Parabola

A key to saving time in the manufacturing shop, is to be able to inspect the mirror while it remains in the fixture. One approach that is being pursued is a modified scanning Hartmann technique. The device, shown in Figure 8, has been built by Talandic Research Corporation. While the hardware is in the exploratory phase of its deployment, it has attractive potential. Those advantages are that it is non-interferometric with reduced sensitivity to vibration and thermal gradients and can test aspherics as well as spheres without auxiliary optics. Also very fast optics can be tested and readings can easily be integrated into SPT numeric control. A disadvantage is that the basic measurement is slope and those measurements must be properly integrated to determine contour. To date, the technique has not been used in-process but has been demonstrated for on-process. Its capability at this phase of its development is the ability to measure figure accuracy to about one quarter of a wave. Other techniques for on-process figure metrology that are being further developed include point polarization interferometry (developed for RADC) and diffraction null corrector. Overall we desire to focus on aspheric testing.

Figure 8: Modified Scanning Hartman Device Built by Talandic Research Corporation

Although an increasing number of materials can be single point turned, ceramics are more easily ground. If conditions can be established to allow grinding in the ductile regime for a particular ceramic or glass, the surface finish is improved dramatically and the mechanical integrity of the brittle material is enhanced because of fewer surface cracks. In the PVTB, a motorized block head spindle (Professional Instruments) has been installed on a Moore Turning machine, Figure 9. Currently, that ductile grinding cell is being used for exploratory tests on silicon carbide. To date, small SiC coupons, 75mm diameter, have been ground to surface finish of 35Å RMS. A photo of that surface is shown in Figure 10. Work by T. Bifano of Boston University on cubic silicon carbide (CVD) indicate that this type of SiC can be ground without creating new surface or subsurface cracks.

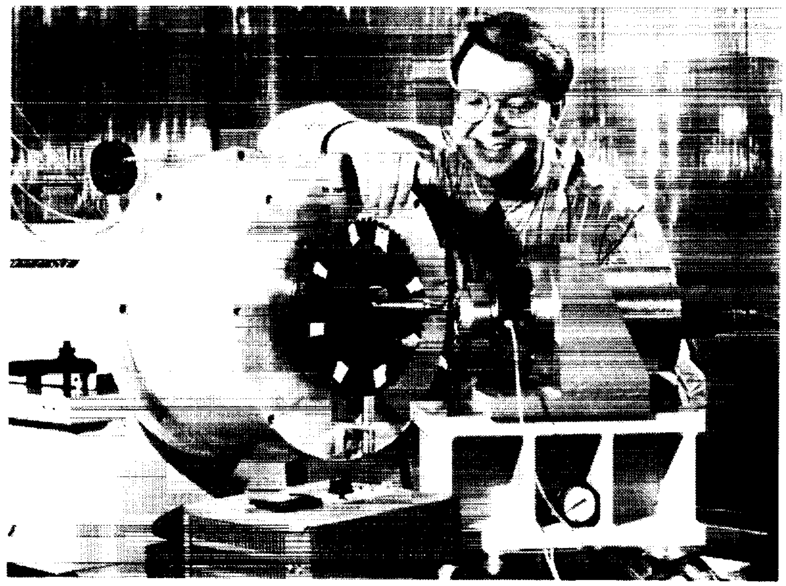


Figure 9: Grinding Spindle Setup in PVTB Cell

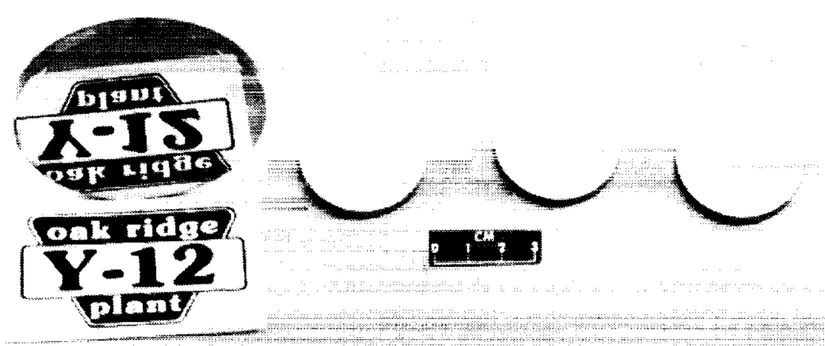


Figure 10: As-ground CVD Silicon Carbide Specimens

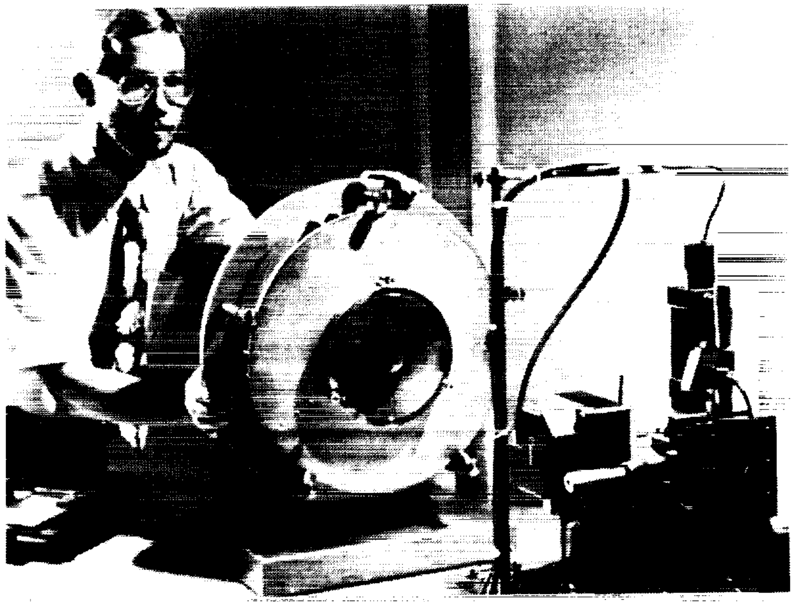


Figure 11: Testing of Optical Figure Using Scanning Hartman Device

Inspection techniques for in/on-process metrology (Figure 11) for grinding of ceramic mirror substrates and for some optical window materials will be the same as used for the Single Point Turning operation. However, acoustic emission techniques are being evaluated to control the grinding parameters within or near the ductile regime. Efforts at Boston University will produce a device to measure acoustic emission caused by cracking of brittle materials during the grinding operation with the potential for real-time feedback as process control. This technique could allow the process to be controlled in the optimum range for grinding.

Beyond the precision capability of ductile grinding and single point diamond turning, is the nanolevel capability of ion milling. The PVTB has two units, one capable of milling a 55mm disc while the second unit is capable of finishing 600mm diameter. The smaller unit is shown in Figure 12. Up to 200 microns of single crystal silicon can be removed by ion milling and the surface finish remains better than about 25 Å RMS. The same is true for non-crystalline materials. However, polycrystalline metals such as aluminum, copper, silicon begin to roughen as the ion milling progresses unless the grain size is very small. To date, removing a layer of about 300 nanometers increased the surface roughness of polysilicon from about 30Å to slightly above 40Å. Tests are continuing. Rates of surface removal for ion milling operations indicate that it would be possible to improve the figure of a mirror from one half wave to one twentieth of a wave in about 3/4 hour for a small optic (100mm) and in about 4 hours for a larger optic (400mm). This assumes that one X-Y raster of the ion gun will be perfect and the in-process inspection technique is fully integrated into the process operation. Currently, our operation is far from that capability, but we strive with our industrial partners to evaluate this technique as a deterministic process because the potential is very high that the manufacturing technology will be ready to be deployed prior to the 21st Century. We are also considering the manufacture of binary optics using these techniques.

Our metrology plans for Ion Milling are different than those we are developing for Ductile Grinding and Single Point Turning. We are examining the potential use of Electronic Holography coupled with interferometry to perform metrology on ion milled parts within the vacuum chamber.

The manufacturing operations and process integration continues for these ultra-precision techniques. While the concept of using as-diamond turned optics for IR application has been proven, there is much work yet to be completed in order to implement these costs and time saving operations in the U.S. industrial base. We applaud the industrial tripartite who are contributing their time, money and other resources to this collective effort. We anticipate that the spin-off to other applications could be numerous and a new manufacturing segment will be born in the United States.

Figure 12: Small Ion-Milling Machine in PVTB

BUSINESS PERSPECTIVES FOR THE FUTURE

If this segment is born, will it survive? As we look about this country, we see marvelous capability to make unique and complex parts in our optics manufacturing base. Our industrial base can make almost anything, given enough time and funding. Many of these expensive facilities were built and utilized for a particular job, but now have a very low utilization factor. On the other end of the ultra-precision scale, we have smaller shops using less expensive equipment, whose operation is labor intensive and also non-deterministic. All of these businesses are generally high technology job shops, but the use factor is not very high.

Given this historical pattern, isn't it likely that the technology being developed by the Optics MODIL and others will either (a) never be incorporated into the optical market place or (b) used to build unique optical components for SDIO and then find no commercial market to sustain it as a viable business.

The notion is proposed that before this country can have an ultra precision manufacturing capability for optics that is commercially viable, one of at least three criteria must be met: (1) The market for ultra precision optics must be significantly larger and more stable, or, (2) Customers must be willing to pay a higher price for optical components so that the industry can generate the resources to support new equipment, etc, or (3) A change in how the manufacturing business is defined for those who make ultra precision optics.

We support the third criteria as the most plausible and one that could be globally competitive. If we can have a mind set that our business is not to manufacture ultra precision optics, but to manufacture ultra precision shapes - what a difference it could make in the potential market size of our business. If the equipment invested in this enterprise is more deterministic, perhaps even "lights-out" holonic operations, and flexible in its capability with regard to materials and shapes, then the manufacturing operations can produce a broad range of products to ultra precise shapes and surfaces.

If a given business redefined its mission to say, we manufacture and sell ultra-precise and complex shapes with a range of surface qualities, i.e., finish and stability, then the Marketing and Customer Service people who support that business become the experts who appreciate the different customer requirements. Those customers are the users of optical components i.e., mirrors, windows, lenses, bearing surfaces, molds, etc.

For the 21st Century, we will have the opportunity to establish an ultra precision manufacturing capability in this country, but a different mind set on the business will have to be developed to make it a viable commercial entity.

ACKNOWLEDGEMENTS

I wish to acknowledge the funding support and guidance provided by G. Stottlemyer of the Strategic Defense Initiative Organization and the excellent work of A. Miller, J. Cunningham, C. Egert, K. Kahl, C. Maxey, P. Steger, and C. Griffies in the development of material removal processes. Among the many contributors to the strategic directions pursued by the Optics MODIL, are the legions of attendees to our Industrial Briefings and Workshops. I thank them all.

N92-22450

INTEGRATED AUTOMATION FOR MANUFACTURING OF ELECTRONIC ASSEMBLIES

T. Joseph Sampite'
CIM Program Manager
Naval Ocean Systems Center
San Diego, CA 92152-5000

ABSTRACT

Since 1985, the Naval Ocean Systems Center has been identifying and developing needed technology for flexible manufacturing of hybrid microelectronic assemblies. Specific projects have been accomplished through contracts with manufacturing companies, equipment suppliers, and joint efforts with other government agencies. The resulting technology has been shared through semi-annual meetings with government, industry, and academic representatives who form an ad hoc advisory panel. More than 70 major technical capabilities have been identified for which new development is needed. Several of these developments have been completed and are being shared with industry.

INTRODUCTION

Since 1985, the Naval Ocean Systems Center (NOSC) has been identifying and developing needed technology for flexible manufacturing (large and small lot sizes) of hybrid microelectronic assemblies. Specific projects have been accomplished through contracts with manufacturing companies, equipment suppliers, and joint efforts with other government agencies.

Since 1986, more than 70 major technical capabilities for which new development is needed have been identified. Under Navy contract, three developments (Distributed Architecture Computer System, Standardized Hybrid Carrier, and Automated Storage Retrieval System) were completed, while three others (Improved Lead Bond Machine, Automated Epoxy Dispensing, and Automated Substrate Attach) have since been developed and marketed by private industry. The resulting technology has been shared through semi-annual meetings with government, industry, and academic representatives who form an ad hoc advisory panel.

During 1988, NOSC worked closely with the Air Force on a joint effort to develop a Universal Robotic Work Cell. Out of that effort came a conveyor mounted robotic handler which featured an alternative version of the Standard Hybrid Carrier and performed both laser scribing of the hybrid and ink jet marking of the sealed package. In a related effort, a prototype of an Automated Diagnostic System was developed. This system consisted of an 8 probe test fixture for hybrid diagnostics, driven by a cell controller which was fully integrated with the factory host computer system.

In 1989, the focus became computer integrated manufacturing. Two initial capabilities were developed. One was a method to direct download a computer aided design file to a pick and place robot and the other was a low cost method to automatically upload and analyze electrical test data. Next, a joint effort was started with the National Institute of Standards and Technology (NIST) to develop a standardized file transfer format. During the next two years, this format will be used to develop generic, open architecture computer aided engineering tools for the automatic exchange of information between design and manufacturing. Other capabilities will be developed to fully use this standard.

The major programs under which these projects were accomplished are called the Integrated Facility for Automated Hybrid Microcircuit Manufacturing (IFAHMM), the Air Force Precision Guided Munitions Industrial Modernization Incentives Program (PGM/IMIP), and the Microelectronics Computer Integrated Manufacturing (MicroCIM) Program.

IFAHMM

In December 1985, a contract was awarded to Teledyne Microelectronics to initiate development of a factory of the future for the manufacture of hybrid microelectronic assemblies. Teledyne is a major manufacturer of hybrid microcircuits for the Department of Defense and is also a major supplier to the medical industry of hybrid components for pacemakers.

Teledyne started their efforts by first modeling the methods currently in use for making hybrids. They then surveyed the industry to determine where the problems were and what technology was available to solve those problems. Their findings were summarized in a list of 15 proposed technology developments which could enable the industry to realize significant enhancements in hybrid manufacturing (1). Those enabling technologies which had the greatest benefit were selected for development.

Distributed Architecture Computer System

At the heart of all modern factories is a centralized computer system with databases and functional controls distributed throughout the plant and connected via an integrated network. Hybrid microelectronics manufacturers have been reluctant to risk profits on the procurement and development of such a computer system which would be not only costly to develop but also costly to maintain. Computer equipment suppliers have been reluctant to spend research and development funds on systems for an uncertain market (2).

Through an agreement reached with IBM, Teledyne served as a beta site for the computer system development. IBM provided development funds and the Navy funded the implementation and engineering support at Teledyne. The result is the Distributed Automation Edition (DAE), currently being marketed by IBM. DAE runs on an IBM AS400 host system and shop floor control is provided by IBM Model 7552 cell controllers. Networking was originally performed using the Manufacturing Automation Protocol used by General Motors but was later replaced with a Token Ring network.

Additional software was developed by the University of California at Los Angeles. This software addressed specific areas of the hybrid facility such as wirebonder control as well as generic capabilities for work in process tracking and automatic storage and retrieval of kitting materials. Teledyne has subsequently modified this for industry release under the product name of "Factory Manager/2".

Standardized Hybrid Carrier

In order to automate handling of the small 2" x 2" hybrid, the hybrid needed to be enclosed by a carrier frame which could be moved from workstation to workstation. The carrier would protect the delicate hybrid from contamination and damage and would be of standard construction so that a minimum amount of reconfiguration would be needed. The carrier material had to be helium non-absorbent for leak testing during final packaging and the surface area of the carrier had to accommodate a bar code.

Teledyne designed a plastic frame that was injection molded around the hybrid case. Inexpensive molds are available up to a 5" x 5" size. A suitable material for leak testing was not found, so the carrier is designed such that the plastic is not exposed to helium during the test.

The carrier has facilitated work in process tracking and eliminated much of the damage previously attributed to handling. However, Teledyne has not yet implemented an automated handling and distribution system.

Automated Storage Retrieval System

Teledyne purchased a carousel storage system from White Carousels, Inc., of Kenilworth NJ, and modified the software to permit integration with a centralized computer system. All storage containers on the carousel are bar coded and all incoming parts are bar coded prior to placement in the storage containers. In the planning department, all parts and materials needed at kitting are listed and entered into the central database. This information is automatically downloaded to the carousel, in accordance with a pre-programmed production schedule. At the carousel, an operator must still remove parts from their containers and place them into a kitting basket. Both the containers and the individual parts are passed by bar code readers for inventory control.

Other IFAHMM Developments

The initial factory modeling at Teledyne was done by BDM International using the Air Force Integrated Computer Aided Manufacturing Definition (IDEF) method (3,4). These models have been reviewed by representatives of hybrid manufacturers at Raytheon, Westinghouse, Hughes, and CTS. While the Information Model (IDEF1) has always proven to be company specific, the Functional Model (IDEF0) has demonstrated a 90% correlation with other hybrid manufacturing facilities and is considered a basic requirement for the determination of production cost drivers (5).

In another effort, Teledyne worked with Teledyne TAC, a supplier of hybrid manufacturing equipment, to develop an automatic substrate attach machine using pre-formed thermoplastic pads as a replacement for epoxy. Although the adhesive is significantly more uniform, failures in shear testing have delayed incorporation into military standards for hybrid manufacturing.

All of the Teledyne/IFAHMM developments were shown in an End-of-Contract demonstration. As of 1990, Teledyne was continuing to demonstrate these products to interested companies. An extensive set of documentation on the Teledyne efforts is available from NOSC. This includes a thoroughly descriptive Final Report (6), from which the foregoing has been extracted.

PGM/IMIP

A development effort at Hughes Aircraft Company, Microelectronics Division, in Newport Beach CA, was undertaken in 1987 with joint funding from the Navy and the Air Force. This was part of an Industrial Modernization Incentives Program and specifically focused on the production line for the Precision Guided Munitions contract. The two primary developments were for a Universal Robotic Work Cell (URWC) and for an Automated Diagnostic System.

Universal Robotic Work Cell

It was originally planned to develop a material handling system which has a universal interface to any shop floor workstation, however, only two interfaces were developed as part of this project. Those interfaces were for the Laser Scriber workstation and the Ink Jet Symbolizer workstation.

The robot is from Intelledex and the software operating system is QNX. A PC is used as the Cell Manager. The Cell Manager interfaces with the company host computer (Stratus), the robot, and each workstation. The architecture of the Cell Manager consists of a Network server, and several Link Drivers.

Material handling is performed by a transport system, consisting of a robotic arm and a barcode reader attached to a carrier basket which itself is mounted on a conveyor belt. The robotic arm picks up cassettes from a loading station located next to the conveyor belt and places them into the carrier basket. The cassette has a barcode which is read as the cassette is placed in the basket. A cassette contains carriers which hold the hybrid substrates.

The carriers are designed to fit most of the standard hybrid sizes. Hughes reviewed the Teledyne design, but considered it not rugged enough for robotic handling. In addition, Hughes desired a carrier which opened up so that the hybrid could be removed, such as during inspection, then replaced afterward back into the carrier. Hughes designed a snap-fit carrier that used composite materials. The material had the extra benefit of not out-gassing helium, so the carrier could be immersed with the hybrid during leak testing.

After loading, the transport system moves down the conveyor to the first workstation, Laser Scribe. There, the arm removes a carrier from the basket and places the carrier into position at the workstation for processing. The workstation secures the carrier and lases a barcode pattern on the substrate. The robot arm then retrieves the substrate carrier, passing the newly scribed barcode pass the barcode reader. When all carriers have been processed, the transport system moves to the next workstation.

Automated Diagnostic System

One of the continuing needs in the manufacture of electronic components is one for automated and integrated diagnostics. Hughes Newport Beach, working with the Micromanipulator Company in Carson City NV, developed a prototype 8-probe system that interfaces with a centralized host computer for both downloading of the parametric and test data, and uploading of the test results. This system allows the rework technician to quickly access data regarding part failure, perform automatically a series of tests to gather specific performance information, then automatically perform the necessary analysis to isolate the cause of the failure.

The interface with the host computer system provides the technician with electronic forms of information, thus eliminating paperwork and the accompanying delays in distributing that paperwork.

Wirebonder Enhancements

Hughes Newport Beach worked in conjunction with Hughes Industrial Products Division in Carlsbad CA, and Gonzaga University in Spokane, to modify the Hughes Model 2460 Wirebonder for direct downloading from a centralized database. Hughes provided a microprocessor interface and operating software was developed by Gonzaga. Accuracy requirements for bonding still necessitate considerable set-up time so the direct downloading feature is not yet cost effective.

MICROCIM

In August 1989, NOSC awarded contracts to the Raytheon Company, in Quincy MA, and to the CTS Corporation, in West Lafayette IN, for the development of methods to implement computer integrated manufacturing technology into the hybrid microelectronics industry. This effort is known as the Microelectronics Computer Integrated Manufacturing (MicroCIM) Program.

The improvements in automated equipment have greatly benefited the electronics industry by increasing throughput with faster performance, increasing yields through consistent high quality performance, and by reducing production costs overall. However, there is rarely any integration of this automated equipment and there is significant labor required for the manual transfer of information between machines. Because of the chance of error being introduced by the manual transfer, additional costs are incurred to check for errors and to make the necessary corrections. The purpose of MicroCIM is to show the benefits of integrating these "islands of automation".

CAD Downloading

In 1990, Raytheon and CTS each developed some preliminary software for system integration. The Raytheon efforts permitted a computer aided design (CAD) file to be downloaded to a computer aided machine (CAM) on the assembly floor. The CAD system used at Raytheon is Mentor-Graphics. The CAM system was a pick and place robot. Hybrid circuit designs are stored in the Mentor-Graphics machine and have file names which carry a "dxf" extension. A conversion program, written by Raytheon and called CADTOPP, translates the "dxf" file into a format called "comma separated value (CSV)". CSV is a generic data format which is easily read either manually or by machine. The CSV file is stored on a diskette which is carried to the pick and place machine and inserted into the robot's front end processor. A Raytheon software program called "SUBWRITE" then translates the CSV file into subroutines recognized by the robot. Setup time has been reduced by about one minute for each die used on the substrate.

Electrical Test Data Collection

CTS developed software to upload test data into a centralized database for analysis. The overall task was straight-forward but was constrained by the conditions that system cost be low, software interfaces must be generic, and commercially available components had to be used. The intent was to facilitate duplication of this capability by small manufacturers.

The result was a test data collection system, with automatic analysis, for a cost of about $100,000. Hewlett Packard test equipment are each connected to a terminal server which, through an ethernet card, access a local area network. An analysis package is continually running on a workstation in engineering and automatically sends a warning when operating controls are about to be exceeded. The limits for operation are easily changed as needed. The terminal servers, ethernet cards, and network software are low cost, commonly available items. The analysis package is for statistical process control and is available from ATA. Manipulation of the analysis software is done using specially designed operator interface screens which are generic to any hybrid microelectronics manufacturing facility. These interfaces were developed using X-Windows and Motif, which are also low cost, commercially available components.

Since all data is stored on electronic media, paperwork is reduced and analysis is now performed by computer. Fault trends are now recognized instantaneously versus the former 30 minutes using manual methods and failure analysis has been reduced from hours to minutes.

CURRENT EFFORTS

There are currently three other projects underway. One is an Automated Substrate Production Cell at Raytheon. Another is a Manufacturing Data Collection and Analysis project being performed by CTS. The third, and most complex, is the CAD to CAM project being developed in association with NIST with subtasking to Raytheon and CTS.

Automated Substrate Production Cell

The Automated Substrate Production Cell is planned as an integration of several pieces of equipment in the substrate fabrication area. All equipment will be under one cell controller which interfaces with the factory host computer system. In the Cell, an automatic loader will place blank substrates on a conveyor belt which will carry the substrates under a laser scriber for bar coding. The conveyor will then carry the blank substrates under a screening machine which will print an electrical trace pattern on the substrates. The substrates next will be carried to a drying furnace and then automatically inspected for defects by a machine vision system.

Production parameters for any part are downloaded from a centralized database to the cell controller. The cell controller will alarm and stop the production run when test parameters in the machine vision system are exceeded. For routine defects, the controller will send the defective parts to rework where optics will automatically be positioned and focused over the defect to assist the technician. During operation, the statistics gathered by the machine vision system are continually analyzed and uploaded to Engineering for production adjustments.

Goals for this effort include reductions in direct labor, scrap material and rework, and increases in cell capacity and yield. As of September, 1991, a preliminary design has been reviewed and accepted.

Manufacturing Data Collection and Analysis

The Electrical Test Data Collection project served to demonstrate the concept for a low cost approach to factory integration. The next step is to expand that concept. CTS is planning to add other test stations onto their data collection and analysis network. The analysis software will be modified to show specific correlation between the types of defects and the probable process steps which caused the defects. A Design of Experiments effort will help determine which parameters at each process are the most effective to control. Personnel from the Advanced Microelectronics Facility at the Naval Avionics Center are assisting in the identification of the process parameters.

CAD to CAM

As part of a broad based effort to standardize the manner in which information is electronically transferred from design to manufacturing, NOSC is working with the National Institute for Standards and Technology (NIST) to develop an Applications Protocol, in accordance with MIL-D-28000 (7), for Hybrid Microelectronic Assemblies. This work incorporates guidance from the international committee on The Standard for the Exchange of Product Model Data (STEP), and the national committees for the Initial Graphics Exchange Specification (IGES) and the Product Data Exchange using STEP (PDES) (8).

The Protocol has three main elements. One is an activity model which describes the processes used to manufacture a hybrid. The second main element is the information model which relates all data needed for design, manufacturing, test, and documentation to the relevant activities. As of October 1991, these two models have been completed and are being reviewed by IGES/PDES national committee. The third element is an interpretive model which specifies the format into which data elements from specific sources can be mapped.

NIST is the lead agency for the Application Protocol. Raytheon has helped to define the data elements required for manufacturing and CTS has defined those required for design.

CTS will further assist by developing translators to implement the interpretive model. One translator will move data out of a CAD machine and into a centralized database. There, the data can be manipulated and used as needed. CTS will also develop a second translator to move data from a centralized database into a CAD machine. The format for the translation will be IGES.

Raytheon will be developing other software tools to assist in the manipulation of the data contained within the centralized database. This will include the ability to download subsets of the data to specific shop floor machines, the ability to upload parametric data from the shop floor machines, the ability to verify translated files for completeness, and the ability for multiple access of a common set of data while still maintaining configuration control of that data.

All of the work performed during 1990, and the plans for future efforts, are contained in the CTS and Raytheon Phase I Final Reports (9,10).

CONCLUSIONS

The developments undertaken by NOSC are intended to enhance the technical capabilities of the American electronics manufacturing industry. While new manufacturing technology is an important first step, dissemination of that technology to industry and implementation by industry are keys to realizing our intent. Therefore, NOSC continues to sponsor ad hoc advisory panel meetings and has a continuing involvement with industry groups such as the IGES/PDES subcommittee on electronics. In addition, NOSC has taken a leadership role in the development of Navy sponsored Industrial Modernization Incentives Programs to help companies implement new technology as it becomes available.

While these described MicroCIM efforts will conclude at the end of 1992, there are many technology challenges that have yet to be addressed. NOSC has assembled a five-year MicroCIM plan to meet these challenges and has submitted it to the Office of the Assistant Secretary of the Navy for review. For further information on that plan, or on any of the projects presented here, contact T. Joseph Sampite', Naval Ocean Systems Center, Code 936, San Diego CA, 92152-5000 or telephone (619) 553-3265.

REFERENCES

(1) Integrated Facility for Automated Hybrid Microcircuit Manufacturing (IFAHMM), Conceptual TO-BE Factory of the Future Strategic Planning Methodology (SPM) Analysis Results; Naval Ocean Systems Center Technical Document 1368, October 1988.

(2) Manufacturing Technology: Cornerstone of a Renewed Defense Industrial Base; Commission on Engineering and Technical Systems, 1987.

(3) Integrated Facility for Automated Hybrid Microcircuit Manufacturing (IFAHMM), IDEF0 AS-IS Model; Naval Ocean Systems Center Technical Document 1365, October 1988.

(4) Integrated Facility for Automated Hybrid Microcircuit Manufacturing (IFAHMM), IDEF1 AS-IS Model; Naval Ocean Systems Center Technical Document 1366, October 1988.

(5) Lessons Learned in Hybrid Microelectronics ManTech Program-to-Program Technology Transfers; Sanders B. Cox, BDM International; MTAG/IMIP '90 Proceedings.

(6) Integrated Facility for Automated Hybrid Microcircuit Manufacturing (IFAHMM), Final Report; Naval Ocean Systems Center Technical Document 1694, December 1989.

(7) MIL-D-28000, Digital Representation for Communication of Product Data: IGES Application Subsets; December 22, 1987.

(8) IGES/PDES Organization Reference Manual, July 1991.

(9) Microelectronics Computer Integrated Manufacturing (MicroCIM) Technology Development Final Report; CTS Corporation; March 14, 1991.

(10) Microelectronics Computer-Aided Manufacturing (MICROCIM), Phase I Final Report; Raytheon Company; April 26, 1991.

N92-22451

THE AIR FORCE MANUFACTURING TECHNOLOGY (MANTECH) TECHNOLOGY TRANSFER METHODOLOGY AS EXEMPLIEIED BY THE RADAR TRANSMIT/RECEIVE MODULE PROGRAM

Tracy Houpt
AF MANTECH Technology Transfer Focal Point
WL/MTX, WPAFB, OH 45433

Margaret Ridgely
Director, MANTECH Technology Transfer Center
Lawrence Associates, Inc.
WL/MTX, WPAFB, OH 45433

ABSTRACT

The Air Force Manufacturing Technology program is involved with the improvement of radar transmit/receive modules for use in active phased array radars for advanced fighter aircraft. Improvements in all areas of manufacture and test of these modules resulting in order of magnitude improvements in the cost of and rate of production will be addressed, as well as the ongoing transfer of this technology to the Navy.

MANUFACTURING TECHNOLOGY BACKGROUND

Since its inception in 1947, the goal of the Air Force Manufacturing Technology (MANTECH) program has been to enhance productivity, increase quality, and reduce life-cycle cost of weapon systems. Contractual projects are application oriented, designed to demonstrate, validate, and implement manufacturing processes for use by the aerospace industry and the Air Logistics Centers of the Air Force Logistics Command.

MANTECH investments address high-payoff problem areas in all industry sectors producing and repairing weapon systems and support equipment for the Air Force (AF). Problems addressed are generic in nature, applicable to virtually all manufacturers in any industry sector and to multiple weapon systems. Efforts address all levels of industry from large prime contractors to material and parts vendors as small as 20-person shops.

MANTECH investments are made to accelerate and broaden the implementation of production concepts and techniques proven feasible in the research and development community. Contracts are awarded to private industry on a competitive basis and provide focus, direction, and "seed money" in manufacturing technology areas that offer potentially high payoff but are beyond the normal risk for industrial investment. High payoff can be measured not only in direct production savings but also in quality which improves safety, serviceability, and readiness. Projects funded by MANTECH generate and disseminate technical information and technical knowledge. Industry, however, is responsible for direct implementation costs and capital equipment procurements.

The Wright Laboratory's Manufacturing Technology Directorate (MT) is organized into four divisions: Electronics, Integration Technology, Processing and Fabrication, and Industrial Base Analysis; and three offices: Concurrent Engineering, Business Integration, and Defense Production Act.

The Electronics Division (MTE) consists of the Components Fabrication and Assembly Branch, which pursues Air Force needs in solid state microwave systems, microwave tubes, infrared detectors and other energy conversion components, and the Materials and Device Processes Branch, which manages programs in semiconductor materials, digital integrated circuits, interconnections, inspections and tests.

The effective integration of processes, systems, and procedures used in the production of aerospace systems using computer technology is managed by the Integration Technology Division (MTI). Under its auspices are the Information Management Branch, which is actively involved with information management, information sciences and integration, and the Implementation Branch, whose technology areas include computer integrated manufacturing, engineering design, operations research, and material handling and assembly. The Integration Technology Division combines design, manufacturing, and supportability functions within the same organization.

The Processing and Fabrication Division (MTP) manages programs to improve structural and nonstructural materials processing and fabrication. Within this division, the Metals Branch directs the manufacturing methods program for metals and metal matrix composites processing and fabrication. The Nonmetals Branch directs the manufacturing methods programs, which include all manufacturing processes for producing and utilizing propellants, plastics, resins, fibers, composites, fluid elastomers, ceramics, glasses, and coatings.

The objective of the Industrial Base Analysis Division (MTA) is to act as focal point for the USAF industrial base program for productivity, responsiveness, and preparedness planning. They coordinate annual Air Force Systems Command and Air Force Logistics Command data into the U.S. Air Force production base analysis, recommend investment strategies for the MANTECH element, and provide industrial base analyses and technical assistance.

The Concurrent Engineering Office (MTR) plans, initiates, coordinates and manages programs addressing Integrated Product Development (IPD) which span a broad spectrum of disciplines, including engineering design, manufacturing, quality assurance, and logistics support. This office is also responsible for managing the Manufacturing Science program for the Directorate which focuses on establishing a science base from which to transition new technologies for further refinement by the Manufacturing Technology programs.

The Business Management Integration Office (MTX) coordinates and consolidates the investment strategy for the Manufacturing Technology Directorate. This office also plans, coordinates, and manages the Repair Technology Program (REPTECH), provides technical guidance in the evaluation of proposed Industrial Modernization Incentives Program (IMIP) projects, and is the manager of MANTECH's technology transfer and benefits tracking programs.

The Defense Production Act Office (MTD) serves as the program office for Air Force Title III programs, which establish or expand domestic production capacity for materials that are considered critical to DOD. Title III accomplished this by providing domestic industry with incentives in the form of purchases and purchase commitments for materials.

The Air Force Manufacturing Technology Directorate has recently undertaken a comprehensive development and implementation effort for an internal technology transfer and benefits tracking program. Technology transfer is government fostering of technologies and processes with the interest of industry adoption. MANTECH projects are inherently structured to address generic problems and utilize a particular weapon system for demonstrating first implementation. This process benefits large DOD prime contractors with significant manufacturing technology enhancements. Substantial improvement in technology transfer without a corresponding increase in the AF investment requires a focus on the entire industry, not the just the primes. Increasing the capabilities of the U.S. generic manufacturing base (Air Force Air Logistics Centers, subcontractors, vendors) will provide the AF with a sizeable return on investment and help maintain and improve the DOD and commercial industry posture and position in the global manufacturing arena. Successful technology transfer is a goal of the MANTECH program.

One example of a program within the Electronics Division with technology transfer potential both within the DOD industrial base and for commercial purposes is the Radar Transmit/Receive (T/R) Module Program.

RADAR TRANSMIT/RECEIVE MODULE PROGRAM

Active element phased array systems utilizing transmit/receive (T/R) modules are considered to be one of the most promising technologies for future ground-based, airborne, and space-based radar applications. Advanced aircraft require active arrays for radar systems for detecting and tracking multiple targets, detecting stealthy aircraft, and for the benefits of improved reliability,lower maintenance costs, and reductions in size and weight. For example, the current F-16 radar system has a 100 hour Mean Time Between Failure (MTBF), while the proposed configuration using T/R modules would have an MTBF of between 80,000 and 100,000 hours. However, the T/R module is the major cost driver for an active phased array system. An average phased array system for an aircraft would require 2000 modules. At a current cost of $8000 per module, the cost of modules for one radar would be in excess of $16 million! Because of the high cost and quantities required for these modules, Air Force program offices developing new aircraft are reluctant to commit to an active element phased array design.

Feasibility and validation T/R microwave modules for many new systems have been built in small prototype quantities or very limited production quantities, and their use has been largely limited to ground-based systems to date. Costs are extraordinarily high as a result of complex designs, the need for precision fabrication, the costs of parts and materials, and the general lack of adequate assembly, test and automation equipment. This program is needed to reduce T/R module costs and demonstrate that the technology is producible for the weapon systems of the future. The two goals of the programs are affordability; a cost goal of $400 has been established, and volume production; 1000 modules/day will be necessary for full scale production of aircraft radar.

The development of new manufacturing technology for radar T/R modules carries the possibility of more than one successful concept. For this reason, two contracts were issued for a 42-month technical effort in April 1989. One contract was awarded to Texas Instruments and Westinghouse in a joint venture. The second contract was awarded to Hughes Aircraft. These manufacturers represent a significant portion of the airborne radar industry, and were the two competing contractors for the F-22 radar subsystem. The effort is divided into four phases.

Phase I of the program involves definition of the baseline module specification. Phase II defines materials and identifies manufacturing issues related to performance, producibility, and cost and proposes a T/R module configuration that meets the baseline requirements. Phase III will produce a module prototype that addresses pertinent design for manufacturability and economic issues. Phase IV will establish the manufacturing processes and controls to demonstrate the production capability for large quantities of low cost T/R modules.

Currently, the two programs have proceeded though Phase I and are involved in Phase II activities. The specific cost drivers within the T/R module production have been identified as the MMIC chip set, test of the completed module, the module housing, automation of the process, and the rework required to meet specifications. Each of the two efforts is exploring a different set of technologies and applications to address these drivers.

The Texas Instruments/Westinghouse effort involves the transmitter and receiver modules packages in separate, hermetically sealed packages. A U.S. vendor supplied metal three-piece housing is used, consisting of a base, ring, and frame/lid. The module is developed on a multilayer thin film substrate, with integrated RF/DC feedthroughs and wire bond interconnects. A combination of epoxy and solder is used for chip attachment; epoxy for the low power portions, and solder for the high power amplifier. Some of the benefits of this approach include reduction in feedthroughs from 4 to 2, elimination of manual placement by use of chip capacitors compatible with automated assembly equipment, use of low cost flex circuit for the bias and logic interface. Test times have already been reduced over 80% per module - a necessity to meet required throughput and decreased cost goals.

The Hughes approach involves a one package GaAs-based transmitter/receiver. Digital control is provided through a separate package. The package uses a Low Temperature Co-fired Ceramic (LTCC) integrated housing and substrate, and a flip chip mounting. The project has already reduced parts count and interconnects by a factor of four, and Hughes is feeding results from their involvement in the DARPA Microwave Millimeter Wave Monolithic Integrated Circuit (MIMIC) program into this effort to drive the cost of GaAs chips down by a factor of 3 to 4.

At the end of Phase IV of this program, in late 1993, at least one version of a T/R module which meets the requirements of advanced radar systems requirements will be configured for high volume manufacturing. The module will exhibit the following manufacturing qualities: lower parts count than previous revisions (from 135 to 32 parts); fewer fabrication process steps (from 98 to 20 steps); producibility (will demonstrate a large quantity module built within a 30-day calendar period). The module cost will be driven down from $8200 to a projection of less than $400 and a reliability of greater than 125,000 hours Mean Time Between Failure (MTBF) is projected.

In addition to the basic process improvements, efforts are also underway to concurrently transfer this technology to ITT, another radar manufacturer, in support of Navy requirements. This technology transfer activity involves identifying and defining specific Air Force developed technologies within the two programs that have the potential application to manufacturing of Navy C-Band T/R modules. Then, through a combination of hands-on instruction, site visits, and implementation assistance from the two Air Force contractors, the necessary technologies will be transferred to ITT's production facility.

Through this program and others like it, the Air Force Manufacturing Technology Directorate expects to continue its long standing track record of providing high return on investment technologies in the production and repair of Air Force weapon systems.

ELECTRONICS

(Session B2/Room B4)

Wednesday December 4, 1991

- **Gallium Arsenide Quantum-Well-Based Far Infrared Array Imaging Radiometer**
- **A Video Event Trigger for High-Frame-Rate, High-Resolution Video Technology**
- **An Electronic Pan/Tilt/Zoom Camera System**
- **Fiber Optic TV Camera Direct**

N92-22452

GALLIUM ARSENIDE QUANTUM WELL-BASED FAR INFRARED ARRAY RADIOMETRIC IMAGER

Kathrine A. Forrest
Photonics Branch
NASA Goddard Space Flight Center
Greenbelt, Maryland 20771

Murzy D. Jhabvala
Solid State Device Development Branch
NASA Goddard Space Flight Center
Greenbelt, Maryland 20771

ABSTRACT

We have built an array-based camera (FIRARI) for thermal imaging (λ = 8 to 12 microns). FIRARI uses a square-format 128 x 128 element array of aluminum gallium arsenide (AlGaAs) quantum well detectors that are indium bump-bonded to a high-capacity silicon multiplexer (6 x 10^7 electrons full well). The quantum well detectors offer good responsivity (D^* = 1 x 10^{10} cm$(Hz)^{1/2}$/W at 60 K) along with high response- and noise-uniformity (U ~ 0.02-0.04), resulting in excellent thermal images without compensation for variation in pixel response. A noise equivalent temperature difference (NEΔT) of 0.02 K at a scene temperature of 290 K was achieved with the array operating at 60K. FIRARI demonstrated that AlGaAs quantum well detector technology can provide large-format arrays with superior performance to mercury cadmium telluride at far less cost.

INTRODUCTION

The thermal infrared, extending from 8 to 12 microns and beyond, is a band of great interest and utility to many of NASA's remote sensing applications. Studies of the atmosphere, ocean, biosphere and lithosphere in this band provide vital data for energy-balance models that can shed light on the issue of global warming, among other things.

There are many situations where the simplicity of a staring detector array is highly desirable, however, the technological challenge of producing high quality infrared detectors in the 8 to 12 micron waveband has made large-format two-dimensional arrays impractical and prohibitively expensive. Imaging in this band has until now mainly been done by mechanically scanning a scene of interest across a single-point detector (typically mercury cadmium telluride (HgCdTe) at 100 K or extrinsic germanium at 4 K), or by pushbroom scanning across a linear array[1]. Unreliability, high 1/f noise, and large spectral non-uniformities from pixel to pixel are among the problems that motivate the search for detector alternatives to HgCdTe for large-format square arrays[2], whereas the very low operating temperatures of extrinsic silicon and germanium (~25 K and 4 K respectively) make their use for long missions impractical.

The advent and development of molecular beam epitaxy (MBE) and metallorganic chemical vapor deposition (MOCVD) have made possible the growth of semiconductors such as aluminum gallium arsenide (AlGaAs) in very thin layers of precisely controlled thickness. Electrons and holes in these layered structures display quantum confinement effects such as discrete subbands in the conduction and valence bands[3]. These subbands have been exploited to make infrared-sensitive devices[4]. This is in contrast to HgCdTe, with a band-gap energy on the order of that of an infrared photon (~0.1 eV) and which absorbs infrared in interband transitions (shown below) between valence and conduction bands. Quantum well infrared detectors can absorb in intersubband transitions when their energy level spacing is of the same order as the infrared photon energy. Either photoconducting or photovoltaic devices can be made: to make a photovoltaic detector, the multiple quantum wells are inserted between p and n-doped GaAs layers.

AlGaAs single element quantum well IR detectors at $\lambda = 10$ microns typically have sensitivities on the order of 10^{10} cm(Hz)$^{1/2}$/W at 77 K [5]. Although this is significantly lower than for HgCdTe ($D^* > 5 \times 10^{10}$ at 10 microns and 100 K), AlGaAs quantum well detectors do not exhibit any significant 1/f noise, and their response is linear over a wide range of photon flux levels. This makes calibration and pixel compensation of large arrays a much simpler matter than for HgCdTe. The excellent quality of the GaAs substrates and high degree of process control during wafer growth result in layer thickness and compositional variations of much less than 1% across a a three-inch wafer. This uniformity means that large-format square arrays of AlGaAs quantum well detectors can easily be fabricated with uncorrected pixel response and noise uniformity of 2 to 5%. Again, this uniformity is superior to what is attainable in HgCdTe arrays by at least a factor of 2. AlGaAs quantum well photoconductors are high-impedance devices (~100 KΩ), dissipating much less heat than HgCdTe photoconducting detectors. Finally, because of the wider band-gap of AlGaAs (1.43 eV versus 0.12 eV for HgCdTe at 10 microns), these devices are much more radiation-hard than HgCdTe, an important consideration for long duration space flight applications and certain military uses.

Further development of AlGaAs quantum well arrays is now being supported by Defense Advanced Research Projects Agency (DARPA) at ATT-Bell Laboratories in Murray Hill, New Jersey, Rockwell Science Center in Thousand Oaks, California and at Martin Marietta in Catonsville, Maryland. Among the aims of this research are to raise the operating temperature, lower the dark current and evaluate the performance of imaging systems using 128x128 or larger AlGaAs quantum well arrays.

RESULTS

As of May 1990 AlGaAs quantum well detectors had not yet been made in large arrays and successfully integrated with silicon multiplexers, and no performance data existed with which performance predictions could be validated. Under the auspices of the Goddard 1991 Director's Discretionary Fund and with additional support from the EOS project, we designed, built and tested an imaging radiometer based on a 128 x 128 element AlGaAs quantum well detector array. Since this was a Far IR Array Radiometric Imager, the acronym FIRARI was adopted. FIRARI was test-flown in a NASA Skyvan over varied terrain near Wallops Flight Facility in Virginia as part of its performance evaluation.

A schematic of FIRARI is found in Figure 1; its salient features are listed below. FIRARI consists of a 128 x 128 element AlGaAs quantum well detector array with response peaked at 9 microns, indium bump-bonded to a high-capacity silicon multiplexer (full well = 6×10^7 electrons). The detector pitch is 60 microns. The array is housed in a continuous-feed liquid-helium dewar with a heater and temperature controller. Custom long focal-length f/2 zinc selenide and germanium optics with diffraction-limited performance are used without a filter to image a scene onto the array; the pass-band of the optics is approximately 8 to 12 microns. Drive and timing electronics for the array were designed and built inhouse; 12-bit digitization of the pixel signals was obtained via a standard A/D board for the MacII. MacII-based data acquisition, display and analysis software was written inhouse.

Instantaneous field of view (IFOV):	7.5 degrees full angle (131 m^2 foot-print @ 3000 ft alt.) 1 mrad/pixel resolution (1.0 m/pixel @ 3000 ft alt.)
ZnSe & Ge f2 fore-optics: (diffraction-limited)	MTF ~ 0.70 @ 10 line pairs/mm Transmission > 90% @ 9 μm

GaAs-AlGaAs quantum well detectors: (ATT-Bell Labs)	60 μm pitch $D^* = 1 \times 10^{10}$ cm $(Hz)^{1/2}/W$ at 60 K 128 x 128 = 16,384 pixels Operating temp. = 50 K **NEΔT = 0.02 K @ 290 K**
High-capacity silicon multiplexer: (Rockwell Science Center)	Full well = 6×10^7 electrons

MacII-based data acquisition (12-bit digitization) and analysis software

Lakeshore dewar with continuous-feed liquid helium

FIRARI performed very well, producing high-quality images with little need for calibration other than subtraction of a reference frame at ambient temperature. Fewer than 2% of the pixels were bad, i.e. either hot or dark. Measurements in the laboratory of array performance indicated an average blackbody $D^* = (1 +/- 0.03) \times 10^{10}$ cm$(Hz)^{1/2}/W$ at 60 K, yielding NEΔT = 0.020 +/- 0.004 K (values are the mean +/- standard deviation). The quantum efficiency averages 0.1%, which is much lower than anticipated; single, edge-illuminated detectors of the same type have quantum efficiencies of about 25%. Poor coupling at normal incidence into the quantum well structure lowers the quantum efficiency of the array; this is a weakness of this type of detector that needs to be addressed in future development. Furthermore, it is necessary to operate the array at 50 K instead of 77 K in order to reduce dark current to an acceptable level. Future work should be aimed at raising the operating temperature to 77 K in order for this detector technology to be strongly competitive with HgCdTe. This NEΔT is largely the result of very good pixel uniformity (the ratio of the standard deviations to the means of D*, NEDT and quantum efficiency are 2 to 4% before correction) which offsets the low quantum efficiency; likely future improvements in this detector technology should lower the NEΔT still further.

POTENTIAL COMMERCIAL APPLICATIONS

There are many practical applications of thermal imaging in atmospheric windows where there is little absorption by molecular water; these windows are from 3 to 5 μm and roughly 8 to 11 μm[1] . Up to the present time the 3 to 5 μm waveband has been heavily utilized mainly because of the limitations of detectors in the 8 to 12 μm band. Most of the radiance emitted by objects near room temperature falls in the 8 to 12 μm band; furthermore the spectral radiant contrast for near room temperature objects occurs there[6]. Hence the detection sensitivity of many thermal instruments can be improved by using detectors sensitive in the 8 to 12 μm band.

Thermal imagery has many applications in addition to NASA's own interests, including: forest-fire detection[7,8], industrial process monitoring[9,10], and non-destructive evaluation[11] to list a few. For instance, NASA Headquarters (Earth Observations Commercialization Applications Program, Office of Commercial Programs) very recently contributed $600,000 to a joint development effort between NASA Ames Research Center and Terra-Mar Resource Information Services to produce a system to provide ground crews with rapid access to information on forest fires gathered by reconnaissance aircraft bearing infrared sensors[8] As part of this effort, NASA has been called upon to improve its infrared sensors. The goal of this three-year program is "to develop and market a commercially viable real-time remote sensing system for monitoring such disasters as fires, oil-spills and floods." Quantum-well detector arrays such as the one used in FIRARI could in the future be a real asset to each of the applications described above: the square format and large number of pixels make detection more sensitive, data acquisition faster and simplify the detection system by eliminating the scan mechanism required for linear arrays and point detectors.

REFERENCES

1. See for example: *Remote Sensing and Image Interpretation*, Second Edition, T. M. Lillesand and R. W. Kiefer, John Wiley and Sons, 1987

2. Proceedings of IRIS Specialty Group on Infrared Detectors, August 1991

3. An excellent general reference on this subject is: *Quantum Phenomena*, S. Datta, Addison-Wesley Modular Series on Solid State Devices, 1989

4. "A new IR detector using electron emission from multiple quantum wells", J. S. Smith et alia, Journal of Vacuum Science and Technology, V B1, pp 376-378, 1983

5. "High sensitivity low dark current 10 μm GaAs quantum well infrared photodetectors", B. F. Levine et alia, Applied Physics Letters, V 56 - No. 9, pp 851 - 853, February 26, 1990

6. See for example: RCA Electro-Optics Handbook, 1974, pp 35 - 44

7. "Early warning detection of forest fire at high resolution by thermal scanner from high-flying aircraft", A. Rainhart, SPIE Proceedings Volume 1467, 1991

8. "Fighting Forest Fires with Remote Sensing", Photonics Spectra, inset p 66, August 1991

9. "Using IR thermography as a manufacturing tool to analyze and repair defects in printed circuit boards", D. K. Fike, SPIE Proceedings Volume 1467, 1991

10. "A cool box finds hot spots", A. Teich, Photonics Spectra, p 99, August 1991

11. "Real-time thermographic battle damage field inspection for rotary wing aircraft", C. G. Pergantis, SPIE Proceedings Volume 1467, 1991

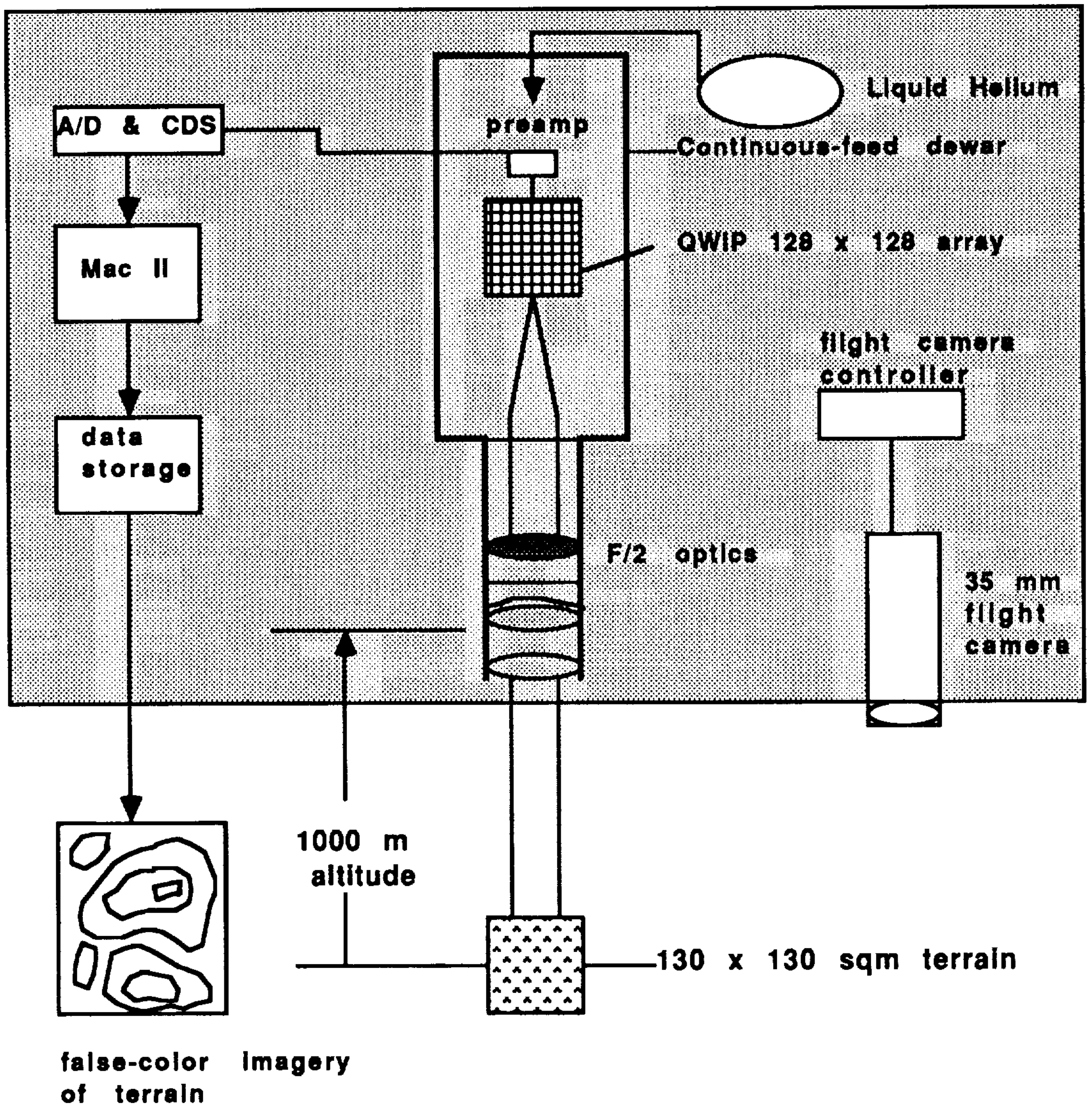


Figure 1. Schematic of FIRARI as flown in NASA Skyvan in vicinity of Wallops Flight Facility, 12:30 to 13:45 EST, May 23, 1991.

N92-22453.1

A VIDEO EVENT TRIGGER FOR HIGH FRAME RATE, HIGH RESOLUTION VIDEO TECHNOLOGY

Glenn L. Williams
NASA Lewis Research Center
Cleveland, OH 44135

ABSTRACT

High resolution, high frame rate video cameras using new image sensor technology will soon be able to make significant contributions in scientific and engineering research and long term scene monitoring. Researchers have in the past relied only on high speed photographic motion picture cameras for recording their scenes, with the attendant loss of real-time access to the images, in order to retain the advantage of high resolution. When video replaces film, while the images are available in real time, the digitized video data accumulates very rapidly, leading to a difficult and costly data storage problem. One solution exists for cases when the video images represent continuously repetitive "static scenes" containing negligible activity, occasionally interrupted by short events of interest. Minutes or hours of redundant video frames can be ignored, and not stored, until activity begins. A new, highly parallel digital state machine generates a digital trigger signal at the onset of a video event. High capacity random access memory storage coupled with newly available fuzzy logic devices permits monitoring a video image stream for long term (DC-like) or short term (AC-like) changes caused by spatial translation, dilation, appearance, disappearance, or color change in a video object. Pretrigger and post-trigger storage techniques are then adaptable to archiving the digital stream from only the significant video images.

INTRODUCTION

Ongoing and future microgravity experiments aboard the Space Shuttle or Space Station Freedom require high resolution high frame rate video technology (HHVT) to replace high speed photographic movie film which is heavy, bulky and which cannot be processed in space [1].

In the 1990's, advances in semiconductor CCD and CID array camera sensor technologies will permit fabrication of high resolution video cameras with frame rates much higher than the commercial broadcast television standard of 30 frames per second. The high resolution of these cameras will mean these cameras will compete with or replace photographic film camera technology. Digitized output from such a high rate video stream presents a difficult data storage problem, because data can be produced at rates 300 Mbyte/sec or higher. When this data is sent to onboard storage, total mission video data storage requirements will easily exceed one terabyte. Without careful attention to cost of storage and transmission, such vast volumes of data will become very expensive to support.

As a means of coping with the potentially huge HHVT data storage requirements, an advanced technology development effort at NASA Lewis has created an architecture for a Video Event Trigger (VET) using digital technology packaged on printed circuit hardware capable of being placed inside a personal computer.

Designed to detect onset of motion within less than 5 milliseconds after a new video frame is available from a digitizer, the system will support acquisition of many seconds of video frame storage when coupled with high density frame store memory capable of continuously recycling uninteresting video frames. With pre-trigger and post-trigger capabilities, such memory could store an entire sequence of images including all the subtle details and changes visible just before the main event.

Built into the VET design is the capability to trigger on sudden image changes while ignoring slow changes, or to trigger on any short term or long term image changes based on differences from stored static reference images.

BACKGROUND

Research and development in the area of detecting and characterizing motion in the video images is not a new idea [2],[3],[4],[5]. Many U.S. and foreign patents cover implementations of scanning and processing video images to filter out noise, locate centroids of motion, and perform data compression of images by coding only areas of images where motion occurs. At the NTSC RS-170 standard video rate of 30 frames per second, hardware exists for quantifying and locating objects in motion within video images. Television scenes from missiles attacking Iraqi targets are recent demonstrations of such technology.

Our need for a Video Event Trigger stems from the HHVT requirements for up to 1000 frames per second of video data, where the ability to trigger on motion within one video frame time would be ideal. In such a case, before the very next video frame is fully digitized, the event would be sensed and special controls would go into effect to begin marking the data for long term storage.

To review, a video camera output drives a coaxial cable with a long sequence of analog signal voltages. Each sequence represents one scan of a raster line across the video scene, and a multiplicity of these sequences, controlled with horizontal and vertical synchronizing pulses, represents one complete video image, or frame. A device called a frame grabber acquires a sequence of raster lines, represents a video frame, and digitizes the signals with an analog-to-digital (A-D) converter, storing the resulting values temporarily in random access memory (RAM). Each 8-bit digital value from the A-D represents a number standing for the brightness and color of a localized region ("dot") of the video image, or "pixel". Screen images are composed of hundreds of thousands of pixels all laid out in columns and rows (such as 512 columns and 480 rows) so that the eye sees a complete picture. At a rate of 30 frames per second, this represents an average data rate of 7.37 million bytes per second. For image sensor devices to be used in the newest cameras, the sensor may actually drive several channels of flash A-D, with the digitized data temporarily going directly to a random access memory.

Of course, the natural consequence of continuously acquiring and digitizing live video becomes a problem of what to do with all the data. At the rate of 30 to 1000 frames per second or more, a real-time frame digitizer would create tens of millions to hundreds of millions of bytes of pixel data per second, to be supplied continuously to high density storage devices, such as magnetic or optical media, or to solid-state RAM. Video RAM or video tape (such as VHS cassettes) or video disk are among the choices of data storage technology in the late 1980's. Note that solid-state RAM storage space represents a high cost limited size memory having nearly zero latency and no moving parts. Whereas, moving magnetic and optical media can store enormous amounts of data permanently, but have quite long latencies. The obvious choice is to take advantage of RAM for temporarily buffering the interesting images during the time necessary to bring the moving media on line.

THE VIDEO EVENT TRIGGER

In typical settings, the human operator in the past has relied on his own observation of the scene or on external devices or externally processed electrical signals, such as from pressure or temperature or acoustic transducers, to create the video event trigger.

A review of processing rates of even the fastest digital processors will reveal that attempts to calculate triggers in software by analyzing frame-to-frame differences will not meet millisecond response requirements. Software algorithms, even on the fastest of computing devices, require many milliseconds up to seconds to analyze the multiplicity of pixels and determine whether the latest video frame has some new "interesting" change happening.

The nature of detecting "interesting" changes is not merely a semantic issue, since changes in color or motion may often be masked by considerable image clutter and may require some algorithmic processing of the image to interpret what is happening. Merely subtracting one video frame from another or looking for motion on the edges of a blob and/or calculating the movement of the centroid of the blob are ways that may fail to generate useful triggers with any one configuration of software rules about what constitutes interesting motion.

By comparison, hardware circuits can be customized to operate faster than software, by embedding comparison algorithms into silicon gate structures special to the application. However, embedded Artificial Intelligence algorithms to process images seeking trigger conditions could yield unsatisfactory results due to an underlying problem in the nature of sensing what is important.

Everyday examples of this problem include:

"Does this frame differ from the last one by more than 5%?"

"The blob didn't move - it changed color!"

"The stripes on that plaid shirt confuse my correlator."

"The cat's iris closed in the bright light, but otherwise stayed in the same place."

The trigger event is some form of change in the picture, determined by any or all of the following changes:

- o Motion of the object of interest or in the area of interest (and ignoring motion elsewhere -- for some scenes may have objects moving elsewhere in the video image at all times, e.g. fans or flashing lights).
- o Change of color in the area of interest.
- o Appearance or disappearance of something in the area of interest.

Some not inconsequential complications to be accommodated are:

- o The movement to be detected may only involve a very small object, which could be lost in pixel noise.
- o Two full frames must be available for comparison. The comparison may not begin until after the end of a frame. The net result is that the trigger process lags one or more frame times behind the most recent frame being acquired.

- o The process of motion interpretation/perception in visual images, which is accomplished in the human brain by millions of neurons making complex rapid-fire decisions, is exceedingly difficult to emulate in circuitry and/or software of reasonable cost and complexity. Software based Artificial Intelligence techniques of the late 1980's are very poorly suited to the task of understanding each and every conceivable video event without major intervention and reprogramming with new computer codes for each task.

One approach to controlling data storage volume and cost is to develop an algorithm which in real-time constrains storage to only include the digitized images of important events. These are defined as images acquired only when there is localized motion around some significant physical event. In NASA's microgravity experiments, video events can often happen after minutes or hours of inactivity prior to the event.

In a subset of cases the video images represent continuously repetitive "static scenes" containing negligible activity, occasionally interrupted by short events of interest. In these cases, minutes to hours of redundant video frames can be ignored until something new happens.

Common examples of such scenes include:

- o automobiles crashing into walls in safety tests, where the front bumper comes into view of the camera only in the last fractions of a second.
- o long term automatic monitoring of parking lots.
- o long term monitoring of entrance to secure or hazardous areas.

For our purposes, we require video event triggering in milliseconds, and simultaneously a low cost, small volume detector. The best compromise appeared to be a fast parallel processor which is autonomous but not elegant. In concert with hardware speedup, there is the need to customize the detection process by adding an ability to make incremental or gross corrections to the algorithm on a frame-by-frame basis.

Fortunately, recent advances in the industry have yielded new techniques and algorithms for performing rapid evaluation of images. These new circuits fall under the collective name of Fuzzy Logic. There are also analog neural networks, which rely on analog components such as capacitors and resistors and operational amplifiers, and digital neural networks, which rely on digital processing of numbers to make similar decisions. Digital neural networks provide extremely consistent decisions without being affected by analog effects such as temperature, charge leakage, and power supply noise. However, to date all neural networks require extensive "training" using "typical" data. But video events are characterized as one-shot unpredictable changes that often are difficult to assign to any particular appearance (as in the above examples).

In our architecture (Figure 1.), a computer containing a "frame grabber" is attached to a video camera, and video frames are acquired using standard hardware setups. Via software and hardware in the computer, control of the acquisition process results in video frames being presented sequentially to the VET logic control circuitry where they are captured and stored into temporary memory buffers (1, 2, ..., 5). Buffer 5 holds the oldest frame, buffer 4 the next oldest frame, and so on, with buffer 1 holding the newly acquired frame, the one to be compared to all the others.

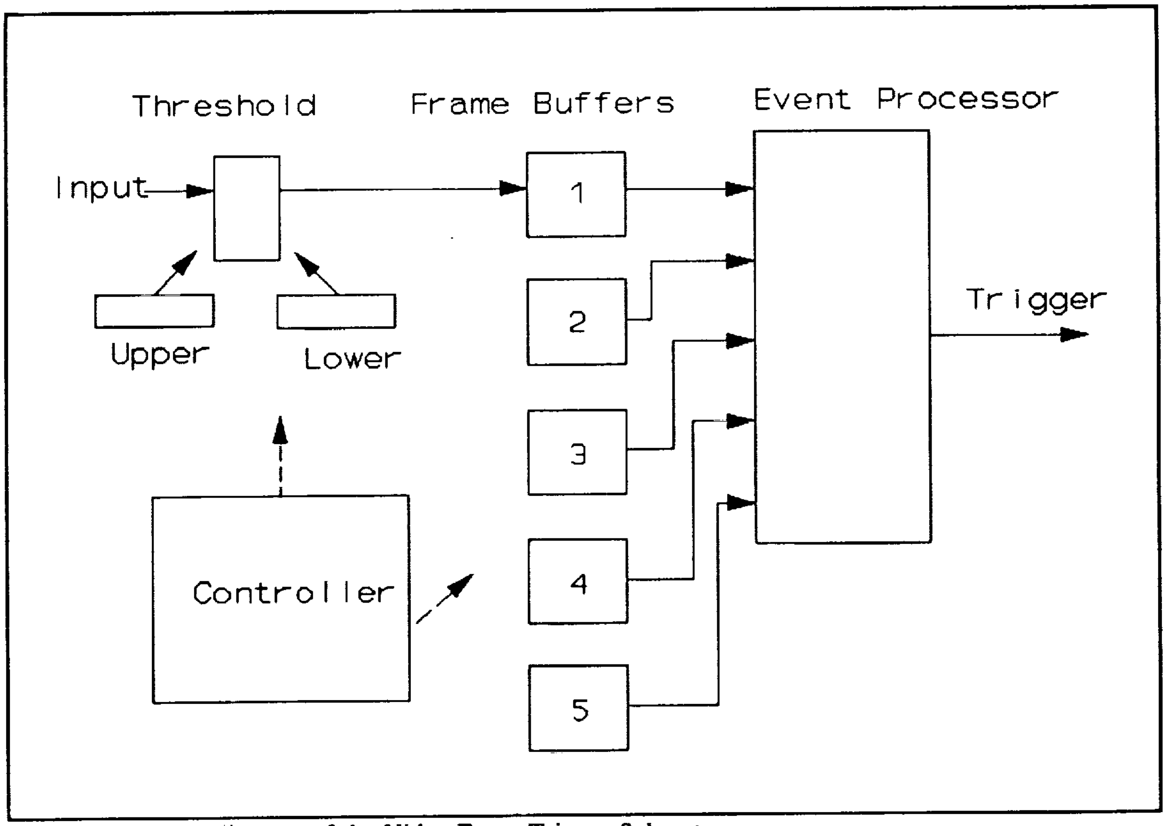


Figure 1 Block diagram of the Video Event Trigger Subsystem.

During system initialization, with the video running continuously, the first few frames are assumed to contain no motion and are stored one-by-one into the k buffers, to be used as the "learned" reference frames. Learning in this case merely means loading up the frame buffer memories, one of which can be loaded every frame time.

There are two methods of storing and comparing old and new video frames. In the first of two methods, each new video frame is compared against a set of k-1 older frames via subtraction and fuzzy logic rules. If no motion is detected, the oldest of the k stored frames is discarded by rearranging pointers to the video frames. Effectively, all the frames are shifted down, and the newest frame is assigned to the first of the reordered k frames. Then the next frame is acquired and the cycle is repeated. This method has the advantage of being able to ignore very slow changes in the video scene, similar to a slow DC drift in a one-dimensional analog signal. Only dramatic changes in the latest video would constitute enough motion to set off a trigger. This is similar in function to "AC Coupling" on an oscilloscope.

In the second method, only the first of the k stored reference frames is ever updated. Each new video frame loaded into buffer 1 is discarded immediately after use. The remaining k-1 stored frames are permanent, non-changing reference frames. The method assumes images can contain either slow changes or rapid changes. ANY changes are important to the motion detection process, and if they occur, they must be reported when a certain threshold is exceeded. This is similar in function to "DC Coupling" on an oscilloscope.

These methods parallel the architectures of one-dimensional FIR and IIR digital waveform filters. However, we avoid the use of recursion, i.e. feeding output images back into the input data stream. We thus avoid problems with instability and limit cycles. But otherwise, our techniques have a close similarity to more commonly digital filter architectures, for which a large pool of documenting literature exists.

We enhance the operation and reduce complexity in the processor by globally thresholding (clipping) the video levels with two digital binary comparators and two "cut" levels. We then have a "window" comparison of the video. "Below Level", "Above Level", "Inside Window", and "Outside Window" are the four choices that result. These levels are merely programmed into storage latches under computer control. These two levels represent the "alpha-cut" (variable sensitivity) levels that determine which levels of gray (or color attributes) will be reduced to a binary ONE by the comparator. All other levels converted to binary ZERO. Then, after the operator's selective adjustment of the alpha-cut levels, the event processor uses only these clipped images. The processing load is thereby greatly simplified, but doing so is not a requirement in the general case, should a design require full gray-level sensing.

ADVANTAGES OF THE VIDEO EVENT TRIGGER

Advantages have already been cited above, in some cases. A summary here would include:

o Able to rapidly sense motion in a video image, in real time operation.

o Relatively inexpensive to construct, and can qualify video images for motion without a large amount of expensive and time consuming hardware or software customizing.

o Ultimately capable of operating stand-alone from any computer and/or microprocessor.

MEMORY REQUIREMENTS

Most applications require minimizing the cost of storage of video images. The quantities of data resulting from high frame rate or image resolution conflict with the need for low cost storage.

In a particular example, assume frames of video data can be stored sequentially and cyclically, a frame at a time, in video RAM storage. High speed, large volume RAM memory boards are commercially available and could in theory be modified for this purpose. Assume the existence of a memory controller, which makes sure that the storage is cyclic, in such a fashion that the very oldest frame is overwritten (lost) each time by a new frame being stored into the memory. We make the "obvious" assumption that the oldest frames carry no data of any value (nothing happened).

Upon an operator "arm" command, the hardware inside the controller starts filling a memory buffer with "pretrigger" data from the digitizer. Once the minimum requirements of the pretrigger buffer have been satisfied, the remaining portion of the buffer is treated as post-trigger data. During the interim, until the trigger signal arrives, the memory is controlled in a fashion similar to a continuous loop magnetic tape recorder. Since the memory has a maximum capacity, the oldest data is continuously replaced with the newest data until the trigger point.

When a video event occurs, the trigger pulse signals the RAM controller to begin a new phase of frame storage algorithms. In this new phase, some of the oldest frames (still redundant) are overwritten by new, interesting frames containing motion. But a selectable number of frames of medium age are retained in memory because they may contain images of precursor activity important to the history leading up to the event. At the trigger time instant, the act of triggering sets a digital logic switch which causes the wrap-around to cease. Thereafter, when the remaining amount of post-trigger memory is filled, the recording stops. The net effect is that the memory holds the entire useful record of the transient, both before and after the trigger point, depending on the size of the "pretrigger memory" setting. All that is required is a little pointer arithmetic to unwrap the data in memory (Figure 2.).

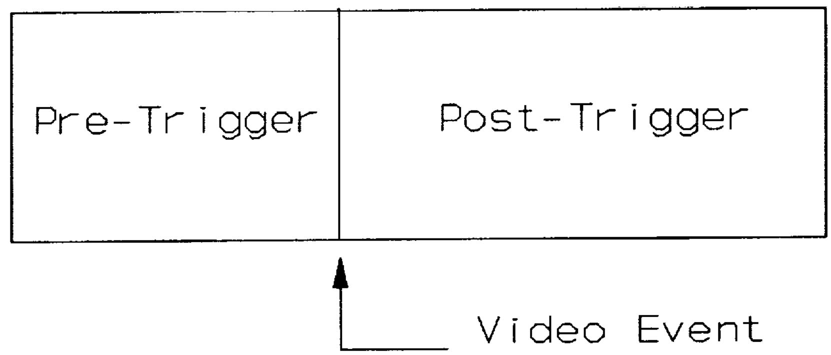


Figure 2 Frame store memory can be divided into pre-trigger and post-trigger portions.

In practice, what the operator directly or indirectly programs into the controller is "pretrigger count" or "pretrigger percentage", the value of which is used to compute a difference. This difference is the total memory size to be used minus the portion to be allocated to pretrigger data. Upon being armed, the controller first blocks and ignores triggers while filling the pretrigger memory portion. As soon as the pretrigger portion is filled, normal triggering is instantly enabled without skipping any frames. Under normal operation, the memory is filled by the wrap-around method described above, with the difference value used to control the size of the post-trigger buffer.

ACKNOWLEDGMENT

This work is being supported by NASA Headquarters Program Office OSSA.

BIBLIOGRAPHY

[1] High Resolution, High Frame Rate Video Technology, *NASA Conference Publication 3080*, May 11-12, 1988.

[2] A. K. Jain, "Image Data Compression: A Review," *Proc. IEEE*, Vol. 69, NO. 3, March 1981, pp. 360-362.

[3] J.R. Jain and A.K. Jain, "Interframe Adaptive Data Compression Techniques For Images," Sig. & Image Proc. Lab., Dep. Elec. and Comput. Eng., Univ. CA, Davis, Aug. 1979.

[4] A. Rosenfeld and A. Kak, *Digital Picture Processing*, Second Edition, Academic Press, 1982.

[5] R. Schalkoff, *Digital Image Processing and Computer Vision*, John Wiley and Sons, Inc., 1989, pp. 190-192.

N92-22454

AN ELECTRONIC PAN/TILT/ZOOM CAMERA SYSTEM†

Steve Zimmermann
TeleRobotics International, Inc.
7325 Oak Ridge Highway
Knoxville, TN 37921

H. Lee Martin
TeleRobotics International, Inc.
7325 Oak Ridge Highway
Knoxville, TN 37921

ABSTRACT

A camera system for omnidirectional image viewing applications that provides pan, tilt, zoom, and rotational orientation within a hemispherical field-of-view utilizing no moving parts has been developed. The imaging device is based on the effect that the image from a fisheye lens, which produces a circular image of an entire hemispherical field-of-view, can be mathematically corrected using high speed electronic circuitry. More specifically, an incoming fisheye image from any image acquisition source is captured in memory of the device, a transformation is performed for the viewing region-of-interest and viewing direction, and a corrected image is output as a video image signal for viewing, recording, or analysis. As a result, this device can accomplish the functions of pan, tilt, rotation, and zoom throughout a hemispherical field-of-view without the need for any mechanical mechanisms. A programmable transformation processor provides flexible control over viewing situations. Multiple images, each with different image magnifications and pan-tilt-rotate parameters, can be obtained from a single camera. The image transformation device can provide corrected images at frame rates compatible with RS-170 standard video equipment. The device can be used for many applications where a conventional mechanical pan-and-tilt orientation mechanism might be considered including inspection, monitoring, surveillance, and target acquisition. Omniview is ideal for multiple target acquisition and image stabilization in military applications due to its multiple image handling and fast response capabilities.

INTRODUCTION

Camera viewing systems are abundant in surveillance, inspection, security, and remote sensing applications. Remote viewing is critical for robotic manipulation tasks, and is often performed by cameras attached to mechanisms that provide the pan, tilt, zoom, and focus capabilities. Close viewing is necessary for detailed manipulation tasks while wide-angle viewing aids positioning of the robotic system to avoid collisions with the work space. The majority of these systems use either a fixed-mounted camera with a limited viewing field, or they utilize mechanical pan-and-tilt platforms and mechanized zoom lenses to orient the camera and magnify its image. These mechanisms can be large, unreliable, and may cause interference and collision with the environment. In the applications where orientation of the camera and magnification of its image are required, the mechanical solution is large and can subtend a significant volume making the viewing system difficult to conceal or use in close quarters. Also, several cameras may be necessary to provide wide-angle viewing or complete coverage of the work space. Camera viewing systems that use internal optics to provide wide viewing angles have been developed in order to minimize the size and volume of the camera and minimize the amount of intrusion into the viewing environment. These systems rely on the movement of either a mirror or prism to change the tilt-angle of orientation and provide mechanical rotation of the entire camera to change the pitch angle of orientation. Using this approach, the size of the camera orientation system can be minimized, but "blind spots" in the center of the view result. Also, these systems typically have no means of magnifying the image and or producing multiple images from a single camera.

In order to minimize the size of the camera and orientation mechanism, a camera system was developed for remote viewing applications that utilizes fisheye optics and electronics processing to provide pan, tilt, zoom,

†Work sponsored by NASA Langley Research Center under contract NAS1-18855.

and rotational capabilities within a hemispherical field-of-view with no moving parts. The Omniview camera approach is based on the property of a fisheye lens which allows a complete hemispherical field-of-view to be captured, but with significant barrel distortion present in the image periphery. A high speed image transformation processor has been developed that reconstitutes portions of the image with the correct perspective for display on an RS-170 standard format monitor. The Omniview camera system has several advantages over other camera systems. The implementation is such that multiple images may be simultaneously produced by the device allowing a single omnidirectional camera to provide numerous views from one location. The transformation is accomplished electronically, providing complete programmable control over viewing parameters. Image sizes, viewing directions, scale and offset etc. may be adjusted to fit operator needs. Since the fisheye image is symmetrical about the image center the camera need not be oriented vertically and can be rotated to match operator.

Potential applications of the Omniview system include remote viewing in constrained environments, inspection, object tracking, and surveillance applications. Since the Omniview camera system can produce multiple images from a single fixed camera, it can replace several camera systems in a remote viewing system.

DEVELOPMENT OF THE MAPPING ALGORITHM

The postulates and equations that follow are based on the camera system utilizing a fisheye lens as the optical element. There are two basic properties and two basic postulates that describe the perfect fisheye lens system. The first property of a fisheye lens is that it encompasses a 2π steradian or hemispherical field-of-view and the image that it produces is a circle. The second property of the lens is that all objects in its field-of-view are in focus, i.e. the perfect fisheye lens has an infinite depth-of-field. In addition to these two main properties, the two important postulates of the fisheye lens system are stated as follows:

Postulate 1: Azimuth angle invariability - For object points that lie in a content plane that is perpendicular to the image plane and passes through the image plane origin, all such points are mapped as image points onto the line of intersection between the image plane and the content plane, i.e. along a radial line. The azimuth angle of the image points is therefore invariant to elevation and object distance changes within the content plane.

Postulate 2: Equidistant Projection Rule - The radial distance, r, from the image plane origin along the azimuth angle containing the projection of the object point is linearly proportional to the zenith angle β, where β is defined as the angle between a perpendicular line through the image plane origin and the line from the image plane origin to the object point. Thus the relationship:

$$r = k\beta \qquad (1)$$

Using these properties and postulates as the foundation of the fisheye lens system, the mathematical transformation for obtaining a perspective corrected image can easily be determined. The picture in Figure 1 shows the coordinate reference frames for the object plane and the image plane. The coordinates u,v describe object points within the object plane. The coordinates x,y,z describe points within the image coordinate frame-of-reference.

The object plane shown in Figure 1 is a typical region-of-interest that we desire to determine the mapping relationship onto the image plane to properly correct the perspective of the object. The direction-of-view vector, **DOV**[x,y,z], determines the zenith and azimuth angles for mapping the object plane, **UV**, onto the image plane, **XY**. The object plane is defined to be perpendicular to the vector **DOV**[x,y,z].

The location of the origin of the object plane in terms of the image plane coordinates is given by:

$$\begin{aligned} x &= D \sin\beta \cos\partial \\ y &= D \sin\beta \sin\partial \\ z &= D \cos\beta \end{aligned} \qquad (2)$$

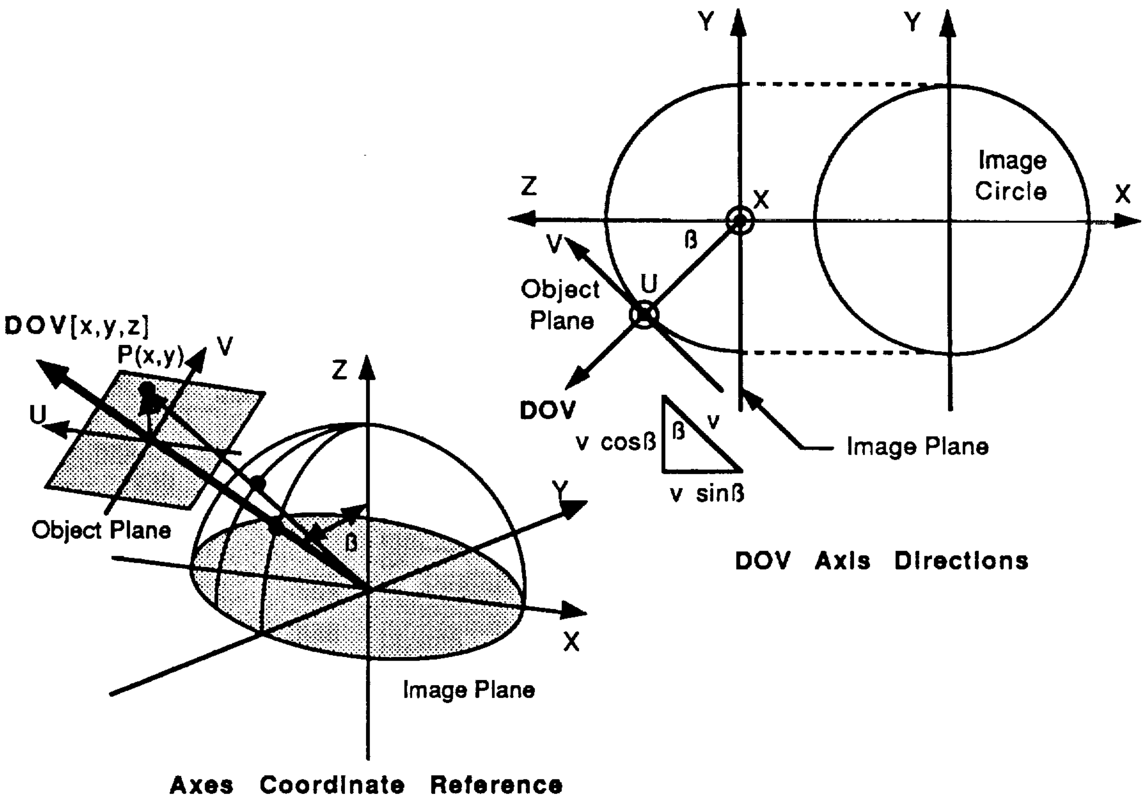


Figure 1 - Coordinate Reference Frame Representation

where D is the scalar length from the image plane origin to the object plane origin, β is the zenith angle, and ∂ is the azimuth angle in image plane spherical coordinates. The origin of the object plane is represented as a vector using the components given in equation 1 as:

$$\mathrm{DOV}[x,y,z] = [\ D\sin\beta\cos\partial,\ D\sin\beta\sin\partial,\ D\cos\beta\ ] \tag{3}$$

DOV[x,y,z] is perpendicular to the object plane and its scalar magnitude D provides the distance to the object plane. By aligning the **YZ** plane with the direction of action of **D**OV[x,y,z], the azimuth angle ∂ becomes either 90 or 270 degrees and therefore the x component becomes 0 resulting in the DOV[x,y,z] coordinates:

$$\mathrm{DOV}[x,y,z] = [\ 0,\ -D\sin\beta,\ D\cos\beta\ ] \tag{4}$$

Referring now to Figure 1, the object point relative to the UV plane origin in coordinates relative to the origin of the image plane is given by the following:

$$\begin{aligned} x &= u \\ y &= v\cos\beta \\ z &= v\sin\beta \end{aligned} \tag{5}$$

therefore, the coordinates of a point P(u,v) that lies in the object plane can be represented as a vector **P**[x,y,z] in image plane coordinates:

$$\mathbf{P}[x,y,z] = [\ u,\ v\cos\beta,\ v\sin\beta\ ] \tag{6}$$

where **P**[x,y,z] describes the position of the object point in image coordinates relative to the origin of the **UV** plane. The object vector that describes the object point in image coordinates is then given by:

$$O[x,y,z] = \mathbf{DOV}[x,y,z] + \mathbf{P}[x,y,z] \tag{7}$$

$$O[x,y,z] = [u,\ v\cos\beta - D\sin\beta,\ v\sin\beta + D\cos\beta] \tag{8}$$

Projection onto a hemisphere of radius R attached to the image plane is determined by scaling the object vector O[x,y,z] to produce a surface vector S[x,y,z]:

$$S[x,y,z] = \frac{RO[x,y,z]}{|O[x,y,z]|} \tag{9}$$

By substituting for the components of O[x,y,z] the vector S[x,y,z] describing the image point mapping onto the hemisphere becomes:

$$S[x,y,z] = \frac{RO[u,\ (v\cos\beta - D\sin\beta),\ (v\sin\beta + D\cos\beta)]}{\sqrt{u^2 + (v\cos\beta - D\sin\beta)^2 + (v\sin\beta + D\cos\beta)^2}} \tag{10}$$

The denominator in the last equation represents the length or absolute value of the vector O[x,y,z] and can be simplified through algebraic and trigonometric manipulation to give:

$$S[x,y,z] = \frac{RO[u,\ (v\cos\beta - D\sin\beta),\ (v\sin\beta + D\cos\beta)]}{\sqrt{u^2+v^2+D^2}} \tag{11}$$

From equation 11, the mapping onto the two-dimensional image plane can be obtained for both x and y as:

$$x = \frac{Ru}{\sqrt{u^2+v^2+D^2}} \tag{12}$$

$$y = \frac{R(v\cos\beta - D\sin\beta)}{\sqrt{u^2+v^2+D^2}} \tag{13}$$

Additionally, the image plane center to object plane distance D can be represented in terms of the fisheye image circle radius R by the relation:

$$D = mR \tag{14}$$

where m represents the scale factor in radial units R from the image plane origin to the object plane origin. Substituting equation 14 into equations 12 and 13 provides a means for obtaining an effective scaling operation or magnification which can be used to provide an equivalent zoom operation.

$$x = \frac{Ru}{\sqrt{u^2+v^2+m^2R^2}} \tag{15}$$

$$y = \frac{R(v\cos\beta - mR\sin\beta)}{\sqrt{u^2+v^2+m^2R^2}} \tag{16}$$

Using the equations for two-dimensional rotation-of-axes for both the **UV** object plane and the **XY** image plane the last two equations can be further manipulated to provide a more general set of equations that provides for rotation within the image plane and rotation within the object plane.

$$x = \frac{R[\,uA - vB + mR\sin\beta\sin\partial\,]}{\sqrt{u^2+v^2+m^2R^2}} \tag{17}$$

$$y = \frac{R[\,uC - vD - mR\sin\beta\cos\partial\,]}{\sqrt{u^2+v^2+m^2R^2}} \tag{18}$$

where

$$\begin{aligned} A &= (\cos\varnothing\cos\partial - \sin\varnothing\sin\partial\cos\beta) \\ B &= (\sin\varnothing\cos\partial + \cos\varnothing\sin\partial\cos\beta) \\ C &= (\cos\varnothing\sin\partial + \sin\varnothing\cos\partial\cos\beta) \\ D &= (\sin\varnothing\sin\partial - \cos\varnothing\cos\partial\cos\beta) \end{aligned} \tag{19}$$

and where

R = radius of the image circle
β = zenith angle
∂ = Azimuth angle in image plane
$\varnothing$ = Object plane rotation angle (20)
m = Magnification
u,v = object plane coordinates
x,y = image plane coordinates

The two equations expressed in 17 and 18 provide a direct mapping from the **UV** space to the **XY** image space and provide the fundamental mathematical foundation for the omnidirectional viewing system with no moving parts. By knowing the desired zenith, azimuth, and object plane rotation angles and the magnification, the locations of x and y in the input image can be determined. This approach provides a means to transform an image from an input image memory buffer to an output image memory buffer exactly. Also, the fisheye image system is completely symmetrical about the zenith; therefore, the vector assignments and resulting signs of various components can be chosen to reflect the desired orientation of the object plane with respect to the image plane. In addition, these postulates and mathematical equations can be modified for various lens elements as necessary for the desired field-of-view coverage in a given application.

DESCRIPTION OF PROTOTYPE CAMERA SYSTEM

The Omniview camera system electronics was implemented using a single wire-wrapped prototype board. A photograph of the Omniview prototype system is shown in Figure 2. A photograph of the prototype electronics board and enclosure are shown in Figure 3. The prototype electronics board was mounted in a 12 by 14 inch enclosure with a +5V and ±12V power supply. A connector was provided for interfacing to the Videk digital camera, a BNC for the monitor display, a DB-15 connector for remote computer control, and a DB-15 connector for the remote hand-held controller. A Videk Megaplus model camera system was used for the development system. The Videk camera was chosen because it had the highest resolution CCD element for a commercially available camera during the initial phase of the project. The Megaplus camera provides a 1320 by 1024 resolution image at up to five frames per second. The Videk camera provided a good platform to demonstrate the prototype system. The camera uses a Nikon F-mount lens adapter and provides a standard 35-

mm back focal plane distance. The input lens element consisted of a Nikon 8-mm fisheye lens and a Nikon lens reduction element. The reduction element was necessary to reduce the 23-mm diameter circle produced by the fisheye lens to match the 2/3-inch format CCD imager size used in the Videk camera. Since the camera provides only a 10 MHz pixel scan rate output, the operating frequency of the wire-wrap implementation was lowered thereby reducing timing constraints.

A block diagram of the prototype system is shown in Figure 4. The camera input image capture electronics consists of a parallel RS-485 type interface to capture the digital output of the Videk camera. The input image memory buffer consisted of a 2048 by 1024 element video RAM array with 8 bit resolution. The input memory buffer was designed using 16 Texas Instruments TMS44C251 1-Megabit video RAMs organized as 512 by 512 element by 4 bit-planes. The output image memory buffer consists of a 1024 by 512 element array with 8 bit resolution and consists of 4 of the TMS44C251 video RAMs. The output display electronics provides a gray-scale 60 Hz interlaced display for an RS-170 standard display monitor.

The microcomputer and control interface for the prototype consists of a simple N80C196 microcontroller host processor core with 64K bytes of RAM and 64K bytes of EPROM. The N80C196 microcontroller provides a great deal of integration and timing control within a single 68-pin PLCC package. It contains an 8-channel 10-bit A/D converter system, several hardware and software timing units, a high-speed input-output peripheral system for event control, a high-speed serial unit and a high speed arithmetic ALU. The 80C196 microcontroller has been used successfully on several embedded Robotics, equipment control, communications, and multimedia platforms. Originally designed for the automotive environment, it provides an excellent core for embedded designs.

Figure 2 - Photograph of Omniview Camera System

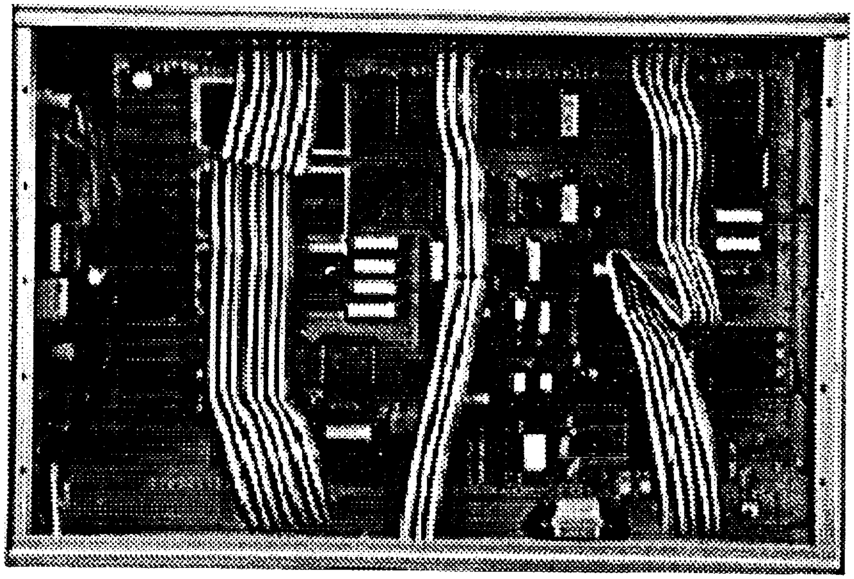


Figure 3 - Photograph of Prototype Electronics Enclosure

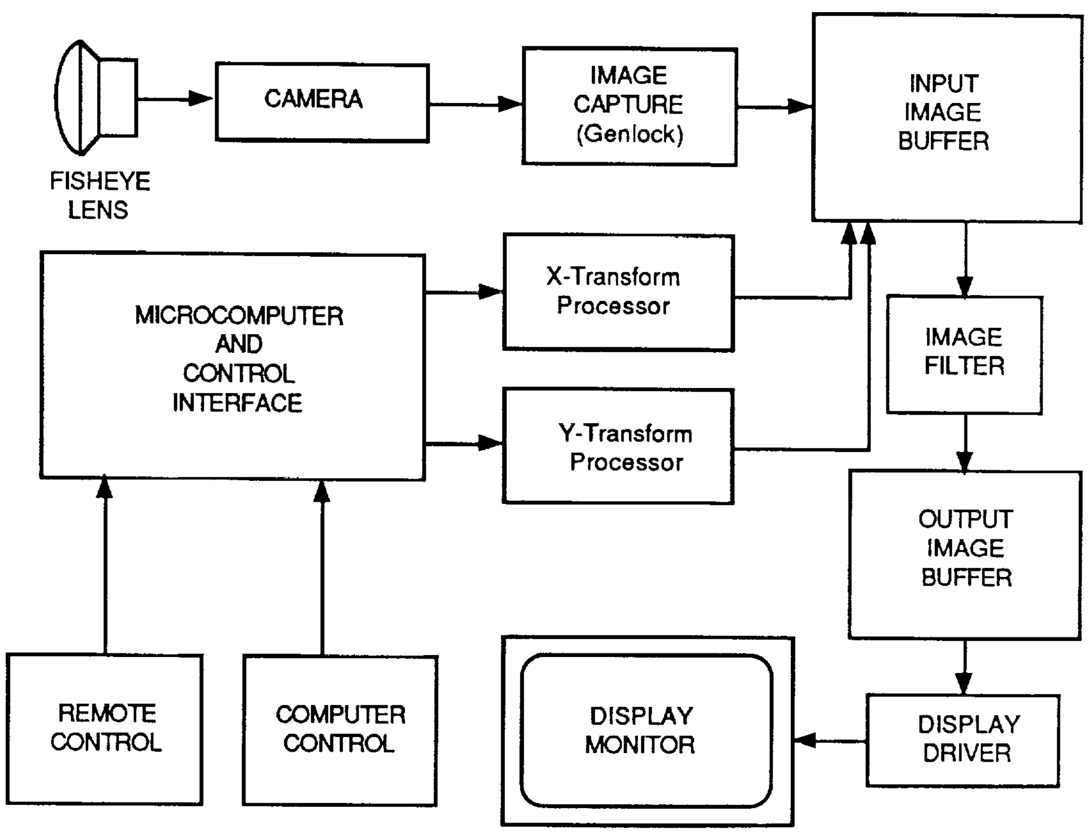


Figure 4 - Omniview Camera System Block Diagram

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

The 80C196 core provides the control interface functions for the prototype system as well as the calculation of the coefficients and parameters for the image transformation core. In order to support the coefficient calculations for the image processor core, a 96-bit and 48-bit arithmetic software package was implemented. The trigonometric functions (sin,cos,tan) were implemented using a lookup table with resolution to within a degree. This was found to be sufficient since the direction-of-view parameters are input to the camera system as direct angles for pan, tilt, and rotation. Also since the frame rate of the Videk camera was only 5 Hz, the calculational update rate of the host processor was sufficient to calculate a new set of parameters during each frame reception. A 48-bit precision square-root function was also implemented using Newton's approximation method.

The image transformation core and image filter consists of high-speed arithmetic devices that implement the basic transformation mapping as presented in equations 17 and 18. There are two independent processor channels that calculate the x and y pixel positions corresponding to the mapped u and v coordinates for each direction-of-view. The image transformation processor is pipelined using both high speed arithmetic devices and FPGA elements in order to maximize overall performance. A single programmable logic sequencer such as the Advanced Micro Devices AM29CPL154 handles all initialization and sequencing operations for the image processor system. The transform processor provides the capability to either nearest neighbor sample the input image space or to provide a 4-pixel bilinear filter on the image. A single 8-bit multiply-accumulate or MAC integrated circuit was used to implement the image filter. The image processor can be developed as a stand-alone core for use in other applications. The present design enhancements of the image transformation processor will allow it to be plugged into future designs.

A hand-held remote control interface was designed to provide an operator input device. The hand-held unit provides two x-y joystick interfaces for independent control of either one or two output images. Rotational potentiometers were provided for control of rotation offset and scale for each image. Using the hand-held unit in the dual image display mode, each image can independently be rotated, panned and tilted to the desired viewing angle. Two toggle switches on the front of the unit provide the capability to select between two different lens configurations and between one and two image display modes.

Photographs of Prototype System Output

Several photographs of the output display of the prototype system in operation are shown in figures 5 through 8. The photograph in figure 5 shows a full 180-degree input image that is being captured by the fisheye lens mounted camera system. The output image is being displayed with no correction being applied. The photograph in figure 6 shows a view looking toward the right. The photograph in figure 7 shows a corrected image looking toward the left. Both corrected images are viewed at angles of up to 90 degrees from the line-of-sight of the camera. As can be seen from the images, the perspective has been corrected so that the barrel distortion evident in the input image has been removed. A good indication of the effect is the correction of the floor pattern and support pole in the fisheye image. The fisheye lens tends to produce circles out of straight lines for lines that are more distant from the center. Using this camera system, the viewing range contains the entire hemispherical field-of-view that is presented to the fisheye lens. One may look from the ceiling to the floor, a range covering 180 degrees, without moving the camera.

The photograph shown in figure 8 shows a simultaneous dual-image display using the same direction-of-view parameters for the images in figures 6 and 7. As shown, the Omniview camera system can display more than one simultaneous image from the same camera. Since the image sizes can be reduced to accommodate more than one image in the same pixel display area, the system is still capable of producing real-time updates with multiple images.

Figure 5 - Fisheye Image

Figure 6 - Right View

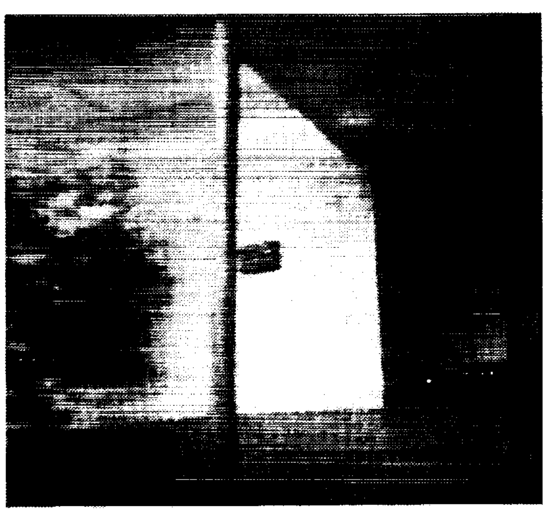


Figure 7 - Left View

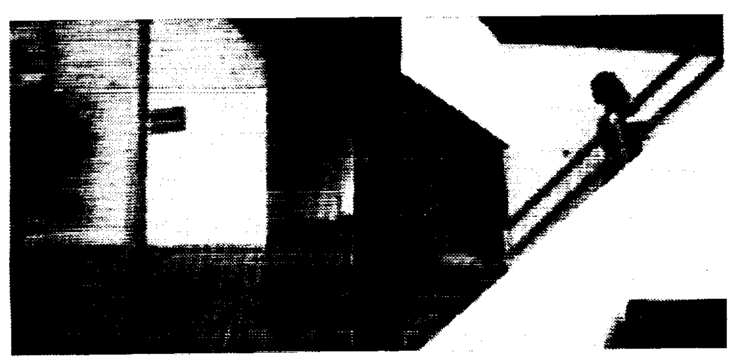


Figure 8 - Dual Left and Right View

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

SUMMARY

The Omniview camera system provides a unique solution for remote viewing requirements. A full hemispherical field-of-view capability provides a wider coverage than most mechanical pan-and-tilt mounted camera systems. The electronic implementation of the pan, tilt, rotation, and zoom functions provides a very flexible system to meet various viewing needs. Slewing from one direction-of-view to another direction-of-view can be accomplished frame-by-frame due to its programmable control of viewing parameters.

A complete prototype camera system has been implemented and demonstrated. The prototype camera system uses a high resolution CCD camera with a Nikon 8-mm fisheye lens and reducer arrangement for the input image capture device. The prototype system provides up to 5 frames per second image acquisition and up to 180-degree field-of-view. Various lens configurations can be used to achieve different field-of-view requirements.

A new design of the electronics is in progress at present. The new design will use an 80960SB 32-bit RISC processor for the host controller. The 80960SB provides a low-cost embedded controller core for high speed mathematical calculations. It contains a full IEEE 80-bit floating point core with support for trigonometric functions. The 80960SB provides enough bandwidth to support multiple image manipulation for a real-time 30 frame-per-second system. The new Omniview electronics system will also support up to 2048 by 2048 element input images for higher resolution viewing applications. The image core of the new design is essentially unchanged except for the integration of the control logic into pipelined FPGA architectures for faster memory access. The video RAM interface logic is being upgraded to support real-time image transformations.

Potential applications of the Omniview system include remote viewing in constrained environments, inspection, object tracking, and surveillance applications. Omniview is ideal for multiple target acquisition and image stabilization applications due to its multiple image handling and fast response capabilities. It can replace several camera systems in a remote viewing application.

The authors would like to express deep appreciation to Walter Hankins, Jack Pennington, and Al Meintel of the NASA Langley Research Center for their support in this research effort.

REFERENCES

[1] S. D. Zimmermann, Notes from Personal Journal.

[2] S. D. Zimmermann, et. al., "Optimizing the Camera and Positioning System for TeleRobotics Worksite Viewing, Final Report for NASA Contract No.NAS1-18627.

N92-22455

FIBER OPTIC TV DIRECT

John E. Kassak
Electronic Engineer
Mail Code DL-ESS-12
John F. Kennedy Space Center, Florida

ABSTRACT

The John F. Kennedy Space Center (KSC) is sponsoring this project to advance the operational television (OTV) technology for the mid 1990's. The objective is to develop a multiple camera system (up to 256 cameras) for KSC installations where camera video, synchronization, control, and status data are transmitted bi-directionally via a single fiber cable at distances in excess of five miles. It will be demonstrated that the benefits [such as improved video performance, immunity from electromagnetic interference (EMI) and radio frequency interference (RFI), elimination of repeater stations, and more system configuration flexibility] can be realized if application of the proven fiber optic transmission concept is used. The control system will marry the lens, pan and tilt, and camera control functions into a modular-based Local Area Network (LAN) control network. Such a system does not exist commercially at present since the Television Broadcast Industry's current practice is to divorce the positional controls from the camera control system. The application software developed for this system will have direct applicability to similar systems in industry using LAN-based control systems.

INTRODUCTION

The fiber optic technology of today can provide tomorrow's systems with improved performance in terms of bandwidth, interference immunity, signal-to-noise ratio, flexibility, and reduction in size, weight, power consumption, and cost. NASA KSC is working on the development of a television camera system that will be the prototype for the next generation of television camera, transmission, and control subsystems for the operational television system to be used at the launch pad. The OTV system at the Kennedy Space Center provides real-time and recorded visual information necessary to conduct and document hazardous and nonhazardous activities during daytime and nighttime operations involving buildup, integration, launch, and landing of the Space Transportation System. This engineering and safety information must be of the highest achievable quality. This quality must be sustained without material degradation during duplication, development of training aids and materials, engineering evaluation, and analysis for the detection, investigation, and correction of anomalies. To that end, each element of the system must meet the maximum performance criteria to ensure optimum overall picture quality and resolution. The goals of the television facility at KSC are to provide to the various users (including the launch team, other NASA centers, and the media) a National Television System Committee (NTSC)[4] and RS-170[2] television signal of the highest quality possible and to act as source material for image analysis, media programming, and launch commitment criteria decision making.

BACKGROUND

The existing OTV system (refer to figure 1) consists of a mixture of color, monochrome, and infrared television cameras. Television cameras with a scan rate approved by the NTSC are presently used to monitor launch operations. Five of the cameras at each pad are broadcast color Charge-Coupled Device (CCD) cameras. The color cameras are strategically located to observe sensitive locations during fueling and launch operations that may be subject to fire. Four combination infrared and visual camera systems are also located around each pad to detect hydrogen fires. The remainder of the cameras are high-resolution, low-light-level monochrome cameras. All cameras are enclosed in hazardproof pressurized housings that provide a controlled environment for the camera and lens. The existing television camera and transmission systems have been developed through an evolutionary process with remnants of the 1960's technology meshing efficiently with today's systems and components. At the launch pad, these cameras are connected to camera control units in the Pad Terminal Connection Room (PTCR) via TV-39 multicore cable. The camera control units contain built-in video equalizers to compensate for loss in transmission on the cable. The video output from the camera control unit is connected to a frequency modulation - radio frequency (FM-RF) modulator. The radio frequency outputs of these modulators are combined onto coaxial trunk cables. The trunk cables proceed through numerous repeater stations to the Launch Control Center. In the Launch Control Center, the radio frequency signal from these trunks passes through a splitter to individual demodulators. This transmission system was designed to meet the requirements of Network Transmission Committee (NTC) Report No. 7[6]. The demodulated video signal is then fed to a video routing switcher for distribution to the end user, to recorders for documentation of launch operations, and to the NASA Select channel for use at other NASA centers. There are 16 pressurized buried RG-247 cables fed from Pad 39A and 16 fed from Pad 39B. Two sync signals are modulated and also transmitted on separate channels from each pad to the Launch Control Center[1].

The FM-RF transmission system RF bandwidth is 6 megahertz (MHz). The modulation bandwidth (BW) is 4.5 megahertz permitting pictures with 356 lines of horizontal resolution to be recorded and viewed by the operational television facilities. The theoretical value for horizontal resolution (Rh) is given by the equation [5]:

$$Rh = \frac{2Ta * \Delta f}{A} = \frac{2(52.6\mu sec.)(4.5MHz)}{1.33} = 355.93\,lines$$

WHERE:
Rh equals the horizontal resolution, in terms of the maximum number of discrete lines observable, when utilizing a test chart, on a horizontal scan line segment whose length is equal to the picture height.
Ta equals the active line scan time in microseconds; for RS-170, TA=52.6 µsec.
Δf equals the available bandwidth of the transmission channel in megahertz; for the FM-RF system, Δf=4.5 MHz.
The A equals the television systems aspect ratio; for NTSC, the picture width is 4/3 of the picture height, so for the NTSC, A=1.33.
(The factor 2 is introduced because each cycle {Hertz} produces an observable picture consisting of a black line and a white line.)

The cameras are controlled at the Launch Control Center operational television control room by a microprocessor-based digital control system, capable of controlling 255 cameras. There are six operating stations in the control room: one master operator and five camera operating positions. There are also maintenance keypads at Pad 39A, Pad 39B, and the Orbiter Processing Facility (OPF). The news facility has a control keypad to control the public affairs color cameras at the pad. Each operating position can control pan and tilt, zoom, and focus on any camera connected to the system. Additional functions available include power on/off, auto or manual iris, and high or normal sensitivity for color cameras. Some cameras carry onboard lights that can be turned on and off by the control system. The existing system was developed by Vicon Industries for monochrome surveillance camera systems. Although the existing camera control system is microprocessor based, the operator control panel and camera interfaces are primarily hardware based. This situation results in system limitations on the type, number, and compatibility of control functions that can be interfaced. A complete hardware design and development is necessary for control of the new system.

PROPOSED SYSTEM CONCEPT

First, the video bandwidth for the transmission from the launch pads to the Launch Control Center must be increased for the higher resolution video available from state-of-the-art cameras. Second, a complete hardware design and development is necessary for the camera control system.

The goal of this development effort is to develop specifications for a multiple color camera system where each camera is interfaced to a single fiber at the pad, to replace the existing system. The opposite end of the fiber is connected to the video camera control system in the Launch Control Center. Refer to figures 2 and 3. The single fiber will utilize bi-directional transmission of video, synchronization, control, and status signals. The concept will eliminate the necessity for camera control and synchronization equipment at the pad and eliminate the requirement for repeaters between the pad and the Launch Control Center. The camera control unit and camera head each contain a fiber optic transmitter and receiver connected to the optic fiber via a wavelength division multiplexer. Signals for synchronization and control of the camera and of the pan and tilt are transmitted from the camera control unit to the camera head on one optical wavelength [1550 nanometers (nm)] while the camera video output and status information are returned on the same fiber utilizing a different wavelength (1300 nanometers). The new camera control system (refer to figure 3) utilizes a host central processor and LAN to provide control, monitoring, and automatic fault reporting for the OTV system.

It is planned to use single mode fiber scheduled to be installed as part of the cable plant in Launch Complex 39 area. At KSC all fiber optic video links are designed to meet the specifications of EIA-RS-250C short haul [3]. The optical fiber cables at KSC contain from 36 to 144 fibers and are installed in existing underground ducts or manholes or are directly buried.

The optical specifications for the KSC single mode fiber are:

1. Equivalent to step index glass
2. Core diameter 8.7 micrometer typical
3. Optical attenuation ≤ 0.5 dB/km (1250 to 1350 nm)
4. Cladding diameter 125 ± 2 micrometer
5. Chromatic dispersion ≤ 0.95 ps/(nm^2-km) dispersion slope at 1310 ± 12 nm wavelength range
6. Mode field diameter of 1300 nm optical spectrum peak within the range of 8.7μm and 9.8μm

Instead of obsolete TV-39 multicore cabling and parallel analog control architecture, the conceptual system utilizes optical fiber and serial data communications to maximize the ability to address the control camera functions, camera types, and other devices. Thus, any camera or device, with a serial control interface, can be integrated into the system permitting full utilization of all the device's functions by remote control.

Each camera station is connected to a transmission system whose only hardware interface requirements are two serial ports and the appropriate video cabling. This allows different types of transmission systems to operate in the system with minimal integration effort. This transmission system need not be used if the camera is close to the control center.

HARDWARE IMPLEMENTATION

The first prototype system was developed and demonstrated in the laboratory before field deployment at Pad 39A. An Ikegami model HC-240 camera utilizing three 1/2-inch frame interline transfer charge coupled devices, with resolution in excess of 700 lines, was used for this phase of the project. The standard camera offered most of the features required for our project including an RS-232 serial data port and published control protocols. This simplified integration with a PC-based control system permitted us to concentrate software development on the user interface and logical groupings for camera, lens, and pan and tilt unit functions. Camera manufacturers do not typically integrate the lens zoom and focus functions or the pan and tilt system control functions into the camera control system. This meant that a separate procurement was required for a pan and tilt system that would support these lens functions and offer a serial data interface. Telemetrics provided the variable speed pan and tilt head, its control system, and the lens integration to support zoom, focus, and remote operation of a 2x lens extender for telephoto applications.

The fiber optic transmission equipment was provided by PCO, Inc. The optical transmission was made on single-mode fiber. Wavelength division multiplexers (WDM) were used at each end of the fiber to permit single-fiber bidirectional transmission. Wavelength division multiplexers allow for two optical frequencies to be inserted on the same fiber. The same device can be used to separate the two wavelengths permitting the use of two independent channels on a single fiber. The analog nature of the signals, the limits of the FM deviation used by the PCO transmission equipment, and the NASA requirements for signal-to-noise ratios dictated the need of high isolation WDMs. Optical crosstalk isolation needed to be 35 decibels or better in order to ensure the signal-to-noise ratios that NASA required after conversion back into the electrical domain. The WDMs selected were manufactured by JDS, Co. and utilize optical filters in their construction. They are quite small and passive and should provide extremely high reliability. The system is designed to operate on both multi-mode and single-mode fibers; however, different WDMs are used for different types of fiber. The system was tested on multi-mode fiber out to 8 kilometers and on single-mode fiber over 17 kilometers with minimal degradation of video quality.

A 1550-nanometer wavelength injection laser diode (ILD) optical transmitter with dual audio channel subcarrier modulators was utilized at the control station end to transmit genlock synchronization video and control data to the camera and pan and tilt systems. At the camera, a PINFET optical receiver and dual audio subcarrier demodulators converted the optical signal back to the electrical domain. A 1330-nanometer ILD transmitter with dual audio subcarrier modulators was utilized at the camera to return the camera and pan and tilt status data along with the camera's video output to the control station.

The control and status serial data were connected via modems to the audio subcarrier modulators/demodulators. The audio subcarrier frequencies were set as near the upper limit of the fiber optic transmitter's video passband as the equipment permitted. This permitted in excess of 8.5 megahertz for the video transmission bandwidth (refer to figure 4). The audio subcarriers were summed with the video channel and this summed signal frequency modulated a carrier which in turn modulated the optical source. At the remote end of the fiber, the PINFET optical receiver detected the optical signal and the resulting FM carrier was demodulated into a video signal and two audio subcarriers. For the video channel, a low pass filter was employed to remove the audio subcarriers from the upper end of the system passband. In an attempt to improve the system video channel group delay performance, the use of notch filters for the audio subcarrier frequencies will be evaluated in the future. The data channels were demodulated by the audio subcarrier demodulators and the outputs connected to modems. The prototype systems performance characteristics can be seen in table 1.

Table 1. System Performance Comparison Table

PARAMETER	PROTOTYPE LAB[1]	PROTOTYPE PAD A[2]	RFA[3]	RFB[4]	EAI-250C SHT. HAUL	NTC-7
Frequency Response (MHz)	.5-8.5	.5-8.5	.5-4.5	.5-4.5	N/A	N/A
(dB)	+0-1.1	+0-1	-.02+.25	-0+.25		
Pulse/Bar Ratio	97.1%	95.4%	102.4%	100.6%	N/A	±6 IRE
2T Pulse K-factor (%)	0.5	0.9	1.4	1.3	N/A	N/A
S/N Unweighted (-) (dB)	61.7	59.1	46.3	46.1	N/A	50
S/N Lum-Weighted (-) (dB)	67.8	66.3	56.8	55.3	67	53
Chroma-Lum Delay (ns)	-3.5	-4.1	-5.6	1.2	N/A	±75
Chroma-Lum Gain	100.2%	95.6%	95.6%	99.5%	±2 IRE	±3 IRE
Differential Gain (%)	0.66	0.84	2.16	3.08	2	15
Differential Phase (degree)	0.22	0.91	1.72	2.60	0.5	5
Lum Nonlinearity (%)	0.37	1.3	4.3	4.9	2	10

1. With a continuous spool of 17.6 kilometers fiber
2. With a fiber path of 17 kilometers with 18 connectors
3. RFA-Radio Frequency Transmission Pad B Channel 7, cable path length 7.2 kilometers
4. RFB-Radio Frequency Transmission Pad B Channel 9, cable path length 7.2 kilometers

SOFTWARE DEVELOPMENT

The initial prototype's control software was developed in C language. The software provides the user access to and status reporting of the camera, lens, and pan and tilt system functions. The user selects the appropriate menu heading of camera, lens, or pan and tilt system and the supported functions are then displayed. When the user selects a function from the menu, a submenu is then displayed to report the system status and the user action required to change the current system status. The initial prototype supports a single camera, lens, and pan and tilt system. The pan and tilt functions supported are the selection of direction (up, down, clockwise, and counterclockwise), the selection of proportional or constant speed, and setting the value for the constant speed and speed range for proportional speed. The lens functions supported are: (1) focus (near or far), (2) zoom (wide angle or telephoto), (3) selection of a 2x extender for telephoto applications, and (4) the selection of the speed that the zoom and focus operations will change. The camera functions supported include: (1) color bars on/off, (2) the selection of inserted title symbology on the camera video (such as a camera number), (3) automatic white and black balance, (4) the selection of the camera shutter speed, (5) the selection of the mode of iris control (manual, automatic, and automatic with manual trim or closed), (6) detail levels, (7) paint controls for individual color channel gain, (8) black level setting, and (9) gain.

A second prototype control system is under development. This system addresses multiple camera systems of differing functions by use of modular software driver packages for each type of system to be controlled. In addition, the software design is object oriented to permit ease of operator reconfiguration as additions, deletions, and changes are made to the system. The second prototype uses a LAN to permit multiple users access to all camera systems on the network as shared resources. Individual permission tables could permit restrictions to be imposed on the functions to be accessed by the user on an individualized basis. The modular-based device drivers permit individual control units to have differing menu functions according to privilege level. These drivers also allow menus to be tailored to address all the specific functions a camera system supports. Many existing camera control systems only offer a hardware-limited subset of the camera system functions. In addition to the control functions, the second prototype system is intended to monitor system status to provide automatic fault reporting. It will also provide remote monitoring and reporting of the pressure and temperature in the pressurized camera, pan and tilt, and light housings.

MEASURED DATA

A series of tests of the existing and prototype television transmission systems was conducted to evaluate the potential benefits of fiber optic transmission for OTV signals used in monitoring and documenting launch processing operations at KSC. These tests supplied both subjective and objective data on the benefits of improvements in this transmission media to the monitoring of launch operations. The prototype system total channel bandwidth of the video and control signals is 12 megahertz. The video bandwidth transmitted is now 8.5 megahertz. Since the existing RF system only allows 4.5 megahertz, the picture resolution is approximately doubled from 356 lines (refer to the previous calculation in the Background paragraph) to 672 lines, where $Rh=[Ta*\Delta f]/A=2[52.6\mu sec.][8.5MHz]/1.33$, since Δf now equals 8.5MHz. Observations made using standard Electronic Industries Association resolution charts support the calculated results. Other measurements are listed in table 1. A decrease of picture distortion is attributable to a differential phase reduction from 2 to 0.91 degree and a differential gain reduction from 3 to 0.84 percent. Other improvements, for example, include an improvement in signal-to-noise ratio from 56 decibels to 66 decibels has been attained under field deployed conditions.

FUTURE DEVELOPMENTS

In 1991, while the first prototype camera station and transmission system is being field deployed to Pad A perimeter site 2, a second prototype camera station and transmission system is being developed. This prototype camera station is intended to be field deployed on the fixed service structure at the pad. Its purpose is to allow the further definition of system specifications in the areas of modular control system design, multiple camera control over a local area network, improved operator interface, automatic monitoring and fault reporting, and definition of other requirements unique to the launch environment related to the new system design. The implications and/or benefits of digital processing cameras and of digital video transmission will also be investigated.

SUMMARY

The application of wavelength division multiplexers for bidirectional fiber optic transmission combined with a modular software-based control system offers an attractive solution to the problems associated with integrating today's high-resolution television camera systems into a demanding operational environment. The technologies required are sufficiently mature and reliable for development into commercial field applications. The main advantages of the system include: (1) the application of a single optical fiber transmission system produces greatly improved video transmission quality and immunity from electromagnetic interface (EMI), RFI, and common-mode problems; (2) no electrical connections are required between remote locations; and (3) conditioning, equalizing, or repeater equipment is no longer required; and (4) a modular software-based control system produces greater flexibility in terms of system configuration, installation, operations, and maintenance. The developed system therefore presents significant advantages in terms of system technical performance and system operations and maintenance.

APPLICATION TRANSFER

The technology developed by this project has benefits to offer in many commercial applications. The developed system is modular in both hardware and software design approach, thus allowing flexibility in initial system design and installation while permitting substantial growth in terms of addressing additional cameras, lenses, pan and tilt heads, recorders, etc., and additional device functions. The developed system is not susceptible to electromagnetic interference or radiofrequency interference. It is weatherproof and hazardproof and offers high quality, secure video transmission in harsh environments at distances approaching 20 kilometers. The potential commercial applications of this technology include: (1) an inexpensive, lightweight replacement for triax-based transmission/control systems for television broadcast cameras in both studio and field applications, (2) wide area coverage for indoor and outdoor closed-circuit television and security systems, (3) high-quality remote visual monitoring of industrial or hazardous processing facilities, (4) monitoring of underwater facilities or operations, (5) remote visual monitoring for offshore oil drilling platforms, and (6) bidirectional intrafacility teleconferencing.

ACKNOWLEDGEMENTS

The author acknowledges the contributions of team members F. H. Galloway and J. E. Brogdon of Boeing Aerospace Operations, Engineering Support Contract and C. D. Smith and D. A. Carter of Lockheed Space Operations Company, Shuttle Processing Contract. The research described was a joint effort between NASA, Boeing, and Lockheed.

REFERENCES

[1] Charles T. Brown, "The Kennedy Space Center Television System," SMPTE Journal, pp 528-540, July 1991.

[2] Electronic Industries Association EIA Standard RS-170 Revision TR 135, November 1957, Electrical Performance Standards-Monochrome Television Studio Facilities.

[3] Electronic Industries Association EIA Standard, RS-250C "Electrical Performance Standards for Television Systems," August 20, 1986.

[4] J. W. Herbstreit and J. Pouliquen, International Standards for Color Television, IEEE Spectrum, March 1967.

[5] Laurence Thorpe, Hideki Miura and Takashi Chikuma, "Large Screen HDTV Monitor Development," SMPTE Journal, pp 620-621, August 1990.

[6] Network Transmission Committee, NTC Report No. 7, "Video Facility Testing Technical Performance Objectives," The Public Broadcasting Service, January 1976.

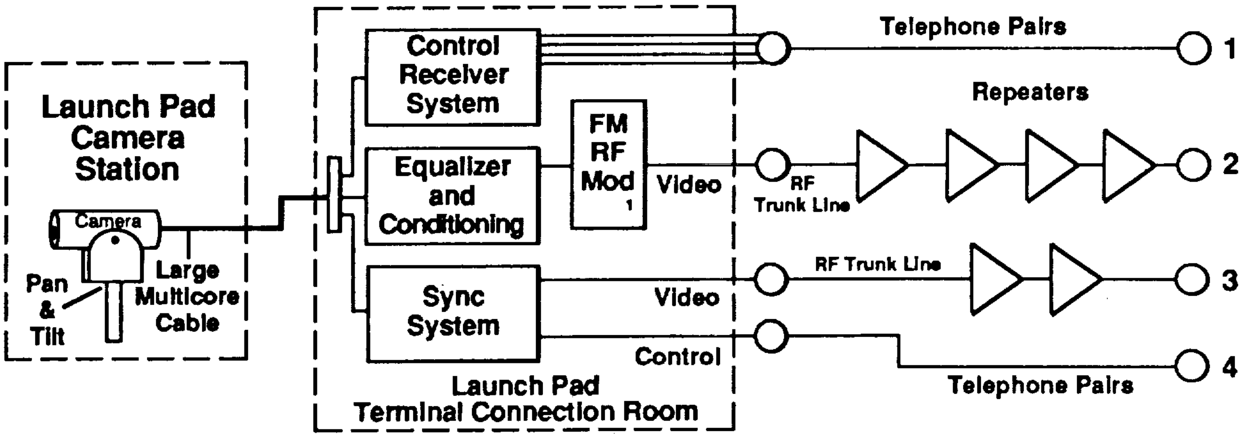


Figure 1. Existing OTV System

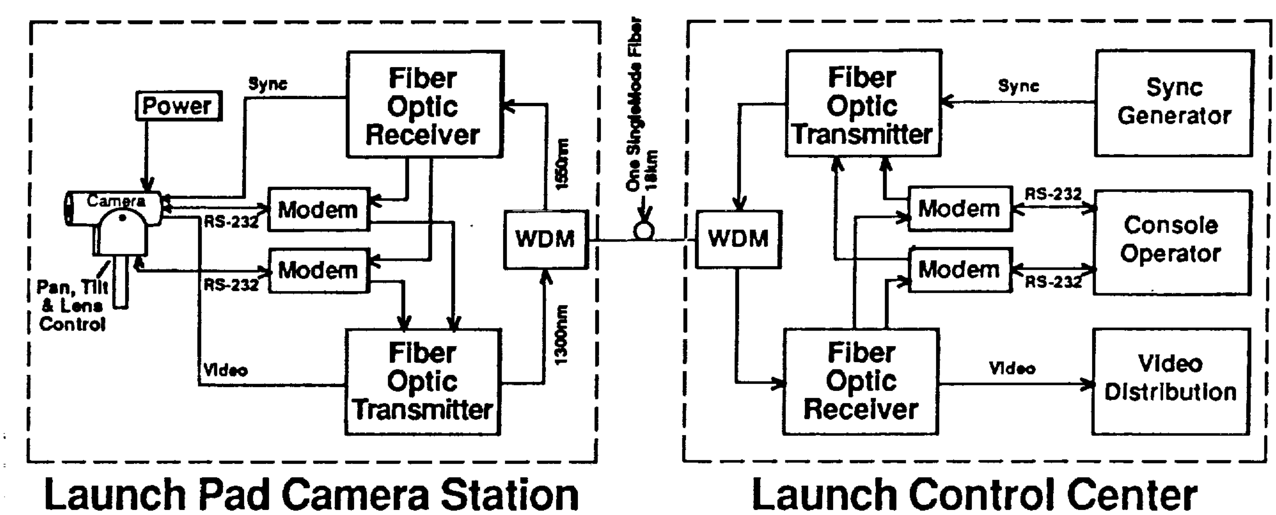


Figure 2. Single Camera OTV System Typical

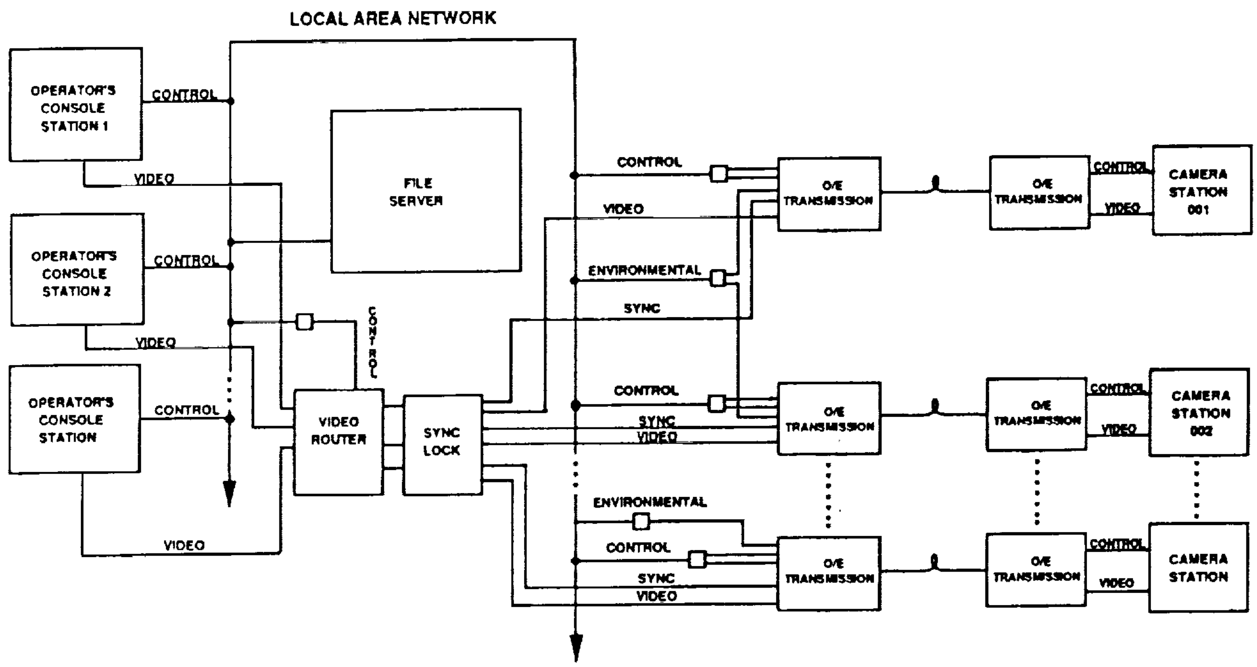


Figure 3. Typical Camera Control System for Multiple Cameras

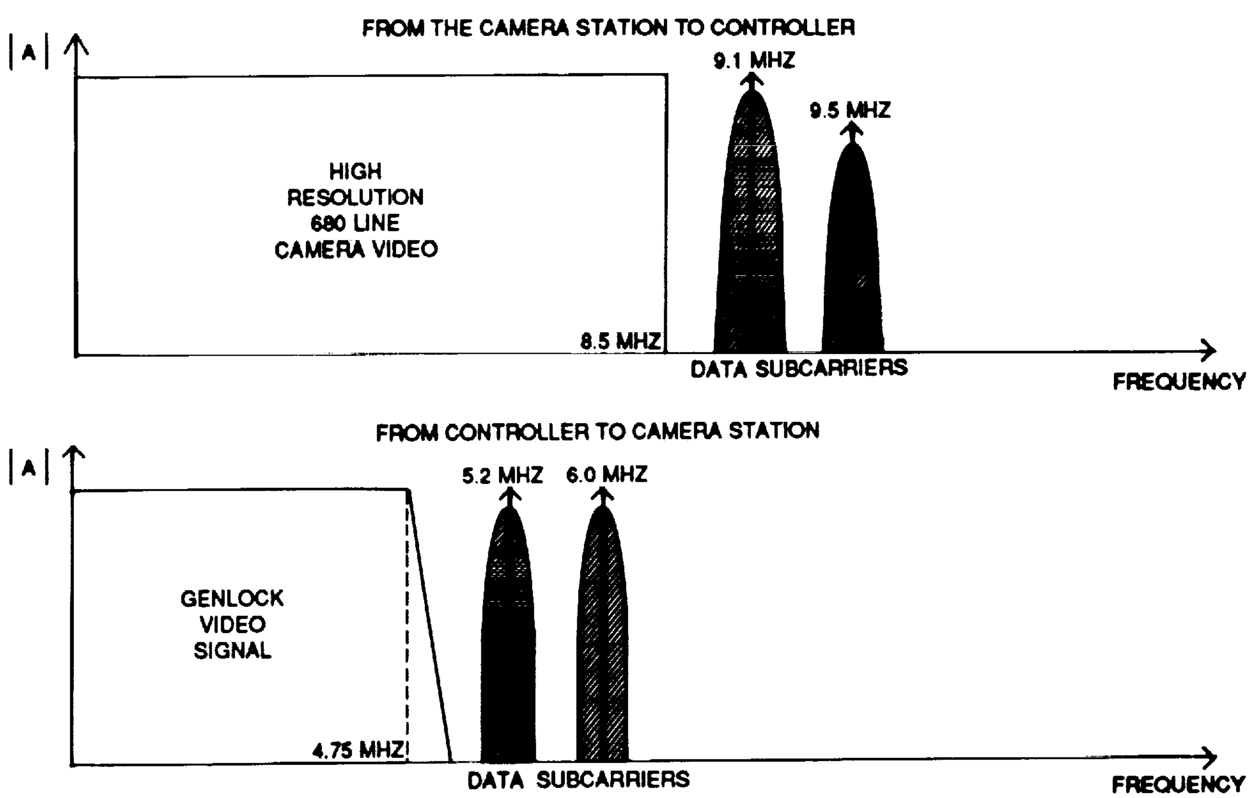


Figure 4. Frequency Modulation Spectrum Designation for Fiber Direct to OTV Composite System

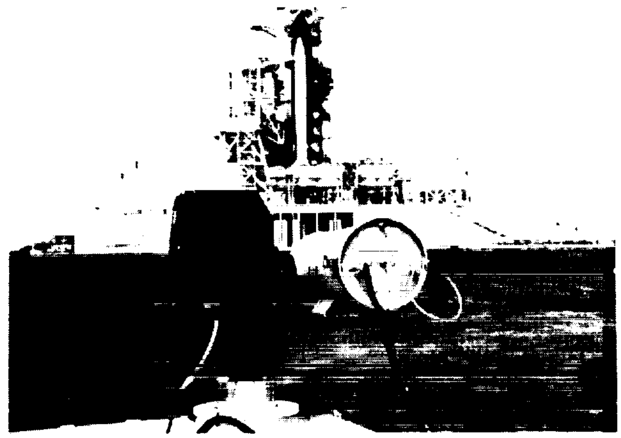


Figure 5. Camera Field Deployment Camera and Pan and Tilt Head

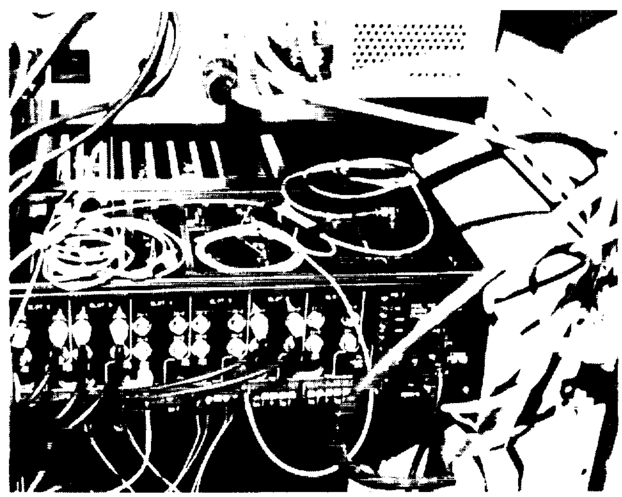


Figure 6. Camera Field Deployment Transmission Equipment

Figure 7. System Control Location Transmission Equipment

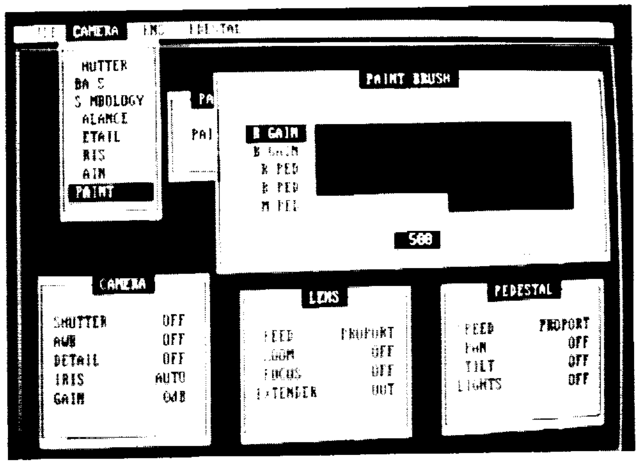


Figure 8. System Operator Interface Camera Menu

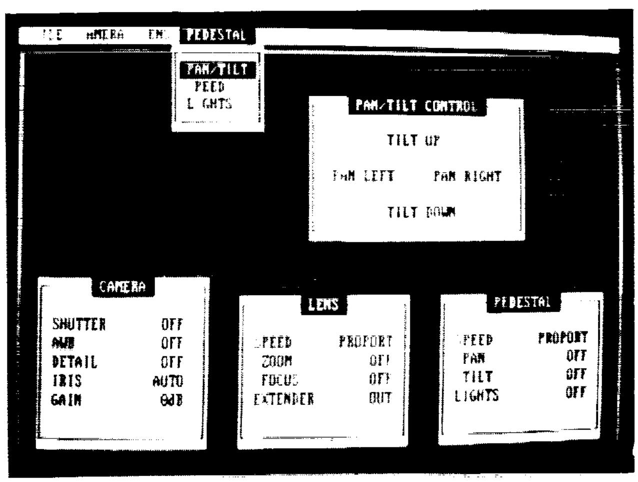


Figure 9. System Operator Interface Pan and Tilt Menu

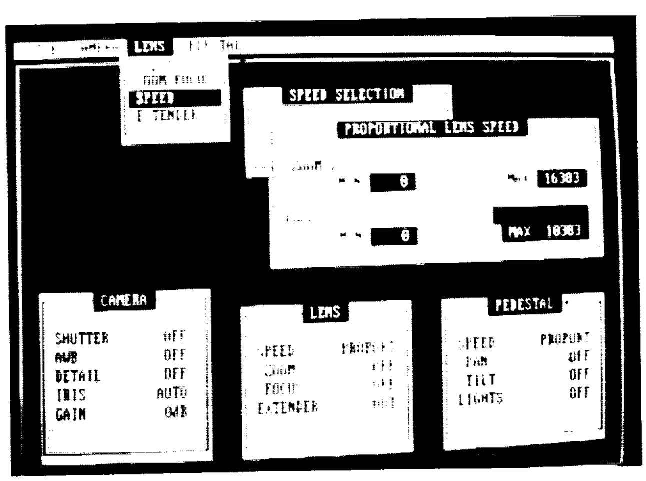


Figure 10. System Operator Interface Lens Menu

ENVIRONMENTAL TECHNOLOGY

(Session B3/Room C1)

Wednesday December 4, 1991

- Waste Management Technology Development and Demonstration Programs
- Regulated Bioluminescence as a Tool for Bioremediation Process Monitoring and Control of Bacterial Cultures
- Fiber-Optic-Based Biosensor
- Ambient Temperature CO Oxidation Catalysts

N92-22456

WASTE MANAGEMENT TECHNOLOGY DEVELOPMENT AND DEMONSTRATION PROGRAMS AT BROOKHAVEN NATIONAL LABORATORY

Paul D. Kalb
Research Engineer
Brookhaven National Laboratory
Upton NY, 11973

Peter Colombo
Technical Program Manager
Brookhaven National Laboratory
Upton NY, 11973

ABSTRACT

Two thermoplastic processes for improved treatment of radioactive, hazardous, and mixed wastes have been developed from bench-scale through technology demonstration: polyethylene encapsulation and modified sulfur cement encapsulation. The steps required to bring technologies from the research and development stage through full-scale implementation are described. Both systems result in durable waste forms that meet current Nuclear Regulatory Commission and Environmental Protection Agency regulatory criteria and provide significant improvements over conventional solidification systems such as hydraulic cement. For example, the polyethylene process can encapsulate up to 70 wt% nitrate salt, compared with a maximum of about 20 wt% for the best hydraulic cement formulation. Modified sulfur cement waste forms containing as much as 43 wt% incinerator fly ash have been formulated, whereas the maximum quantity of this waste in hydraulic cement is 16 wt%.

INTRODUCTION

The Department of Energy (DOE) has generated large volumes of low-level radioactive (LLW), hazardous, and mixed waste as a result of its research and defense activities over the last 50 years. These include a broad range of waste types encompassing diverse chemical and physical properties. The total volume of DOE LLW alone, (either buried or disposed) through 1987 is estimated at 2.4×10^6 cubic meters, and these wastes continue to be generated at an estimated annual rate of 1.5×10^5 cubic meters [1]. Sources of commercial LLW and mixed wastes include nuclear power fuel cycle activities (60%), and industrial/institutional sources (40%) such as hospitals, universities, and radionuclide manufacturers. About 1.5×10^6 cubic meters of commercial LLW has been generated, and the current generation rate is estimated at 5.5×10^4 cubic meters/year. Figure 1 compares current sources of both DOE and commercial LLW.

Much of this waste requires solidification/stabilization (S/S) before final disposal to reduce the mobility of contaminants into the accessible environment. The most common practice at DOE and commercial facilities is to solidify waste using hydraulic cement such as portland cement. Historically, cement processes (also known as grouting) were the first methods for S/S of wastes. They continue to be widely used primarily because they are relatively inexpensive (material costs of $0.10/lb or less), readily available, and easy to process. Cement solidification processes are limited however, because cement hardens by means of a chemical hydration reaction that is susceptible to interferences with the waste. For example, many inorganic salts and heavy metals present in LLW and mixed wastes are known to inhibit cement hydration [2]. These

*This work was sponsored by the U.S. Department of Energy under contract No. DE-AC02-76CH00016.

PRECEDING PAGE BLANK NOT FILMED

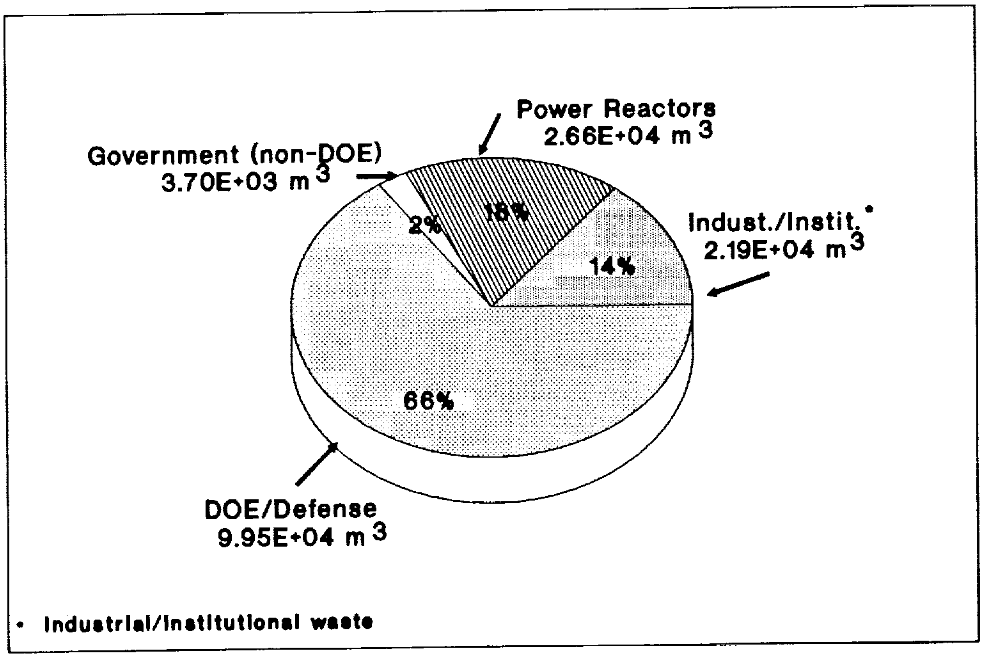


Figure 1 Comparison of DOE and commercial low-level radioactive waste volumes generated annually (1987 data) [1].

interactions reduce the amount of waste that can be incorporated and/or result in poor quality waste forms that do not successfully withstand disposal conditions. Low material costs for hydraulic cement are offset by poor waste loading efficiency (amount of waste per drum) and potentially poor performance of waste forms in disposal. In addition, small variations in waste chemistry over time can necessitate frequent process modifications and quality assurance testing. This seemingly straightforward process can actually prove complex and labor-intensive.

Thus, work conducted at Brookhaven National Laboratory, sponsored by the U.S. Department of Energy Office of Technology Development (DOE OTD), is focusing on the development and demonstration of new and innovative techniques for encapsulation of "problem" wastes. In keeping with OTD's policy of fostering "faster, better, cheaper, and safer technologies," the objectives of this effort are to develop materials and processes that:

- have the potential to encapsulate "problem" LLW and mixed wastes where current practices are inadequate,
- minimize the potential for release of toxic materials to the environment,
- result in durable waste forms that can withstand anticipated conditions during storage, transportation, and disposal,
- are simple to operate, easy to maintain, and economical.

Process development efforts at BNL begin with bench-scale research and development that progresses in logical stages:

1) Waste characterization—Waste composition is analyzed and physical properties relevant to treatment and processing are identified.

2) Investigation of Potential Materials and Processes—A large field of potential systems is examined and the list of preferred candidates is narrowed based on known properties and behavior.

3) Waste-Specific Treatability Studies—Feasibility is investigated based on compatibility of the waste and binder. Preliminary formulations are developed and tested according to broad acceptance criteria (e.g., process results in a monolithic solid waste form with minimum strength characteristics).

4) Formulation Development—The most promising candidates progress to this next phase of R & D in which waste/binder ratios are optimized to provide the best combination of loading efficiency and waste form performance (see below). Often, these two factors represent a trade-off and final formulations must reflect a balance between meeting minimum performance criteria and producing an economical process.

5) Performance Testing—Optimized waste form formulations are subjected to a complete set of waste form property and performance tests to provide a means of comparison among potential S/S options. Performance testing also provides necessary data for predicting long-term behavior of waste forms in storage and disposal. These tests can reveal areas where potential performance improvements are achievable by means of additional waste treatment, modifying formulations, or use of specific additives. In these cases, performance testing and formulation development comprise an iterative process.

6) Economic Feasibility—Overall system cost-effectiveness is examined and compared with conventional and alternative technologies.

Bench-scale systems shown to have potential technical and economic benefits are candidates for process scale-up and technology demonstration activities. The ultimate goal of full-scale technology demonstration is accomplished through the following steps:

1) Site Selection—Appropriate waste generating sites are selected based on types and volumes of waste generated. Cooperation of site personnel is solicited.

2) Feasibility Assessment—Scale-up feasibility is confirmed by means of pilot- or full-scale testing using simulated wastes. Resulting process data are compared with laboratory data and engineering estimates to corroborate scale-up of process parameters. Quality assurance testing of products is conducted to verify proper metering, mixing, and overall product performance.

3) Equipment Acquisition and Installation—Site-specific needs are considered prior to final equipment specification and acquisition.

4) Technology Demonstration—Upon completion of installation, start-up, calibration, and preliminary testing, the technology demonstration is held under actual plant conditions. To provide maximum impact, personnel from throughout the DOE complex, related regulatory agencies, and from the commercial sector are invited to attend.

5) Process Evaluation—Data collected during the technology demonstration is then reviewed to ascertain compliance with quality assurance (QA) requirements and to compare with bench-scale data. Waste form properties of the resulting product are also tested against QA and performance criteria.

6) Technology Transfer—All necessary information for the successful implementation of developed technologies is transferred to target sites within the DOE complex.

SELECTION OF BINDER MATERIALS

Research and development efforts have encompassed waste streams that are common within DOE (e.g., nitrate salts), the commercial sector (e.g., evaporator concentrates, ion exchange resins) or both (e.g., sludges, incinerator ash). Potential S/S binder materials surveyed are listed in Table 1. Because of the inherent problems with using hydraulic cement for solidifying and stabilizing waste discussed above, two thermoplastic materials were selected for further development: low-density polyethylene and modified sulfur cement. These materials can be melted, mixed with waste to form a homogenous mixture and then allowed to cool, resulting in a monolithic solid waste form. Contaminants are immobilized by micro-encapsulation. Since no chemical reaction is required for solidifying, they are compatible with a wider range of waste types and can encapsulate more waste per drum than conventional processes. Process temperatures are relatively low (melting temperature of both materials is ~ 120°C) so volatilization of contaminants is negligible.

Table 1 Potential Encapsulation Materials

Cements	**Thermoplastic**
Portland	Bitumen
Masonry Cement	Polystyrene
Cement-Sodium Silicate	Polypropylene
Pozzolanic	Polyethylene
High Alumina	Modified Sulfur Cement
Blast-Furnace Slag	
Polymer Modified Gypsum	
Polymer Impregnated Concrete	
Glass	**Thermosetting**
Soda-Lime	Vinyl-Ester Styrene
Phosphate	Polyester Styrene
Slag	Water Extendable Polyester

Polyethylene is an organic polymer of crystalline-amorphous structure available in a wide range of densities and molecular weights. These properties, in turn, affect basic material properties such as melt temperature, viscosity, hardness, and permeability. Polyethylene has been shown to withstand conditions that may be encountered in disposal including harsh chemicals, radiation, microbial degradation, freeze-thaw cycling, and saturated conditions [3]. Modified sulfur cement was developed by the U.S. Bureau of Mines about 20 years ago as a means of utilizing by-product sulfur for construction [4]. The supply of sulfur by-products from flue gas de-sulfurization and petroleum refining is growing. More than 5 x 10^6 tons of waste sulfur are projected to be produced annually by the year 2000. Modified sulfur cement is made by reacting elemental sulfur with organic modifiers that increase its stability by suppressing unstable phase transformations. Results of testing have shown that modified sulfur cement is also durable under anticipated disposal conditions [5].

PROCESS DEVELOPMENT AND PERFORMANCE TESTING

Formulation and process development studies have been performed for polyethylene and modified sulfur cement encapsulation of a wide range of waste types. These studies determine the maximum waste loadings for each binder, while still maintaining adequate waste form performance [6,7,8]. Existing waste form performance criteria established by the Nuclear Regulatory Commission (NRC) for demonstrating long-term durability of commercial low-level radioactive waste forms and by the Environmental Protection Agency (EPA) for the toxic leachability of hazardous wastes were applied. Maximum waste loadings achieved for polyethylene and hydraulic cement encapsulation are compared in Figures 2 and 3 (weight % and volume, respectively). Figures 4 and 5 present similar data for modified sulfur cement. Significant improvements in waste loading are attained in each case using these thermoplastic binders when compared with conventional cement systems.

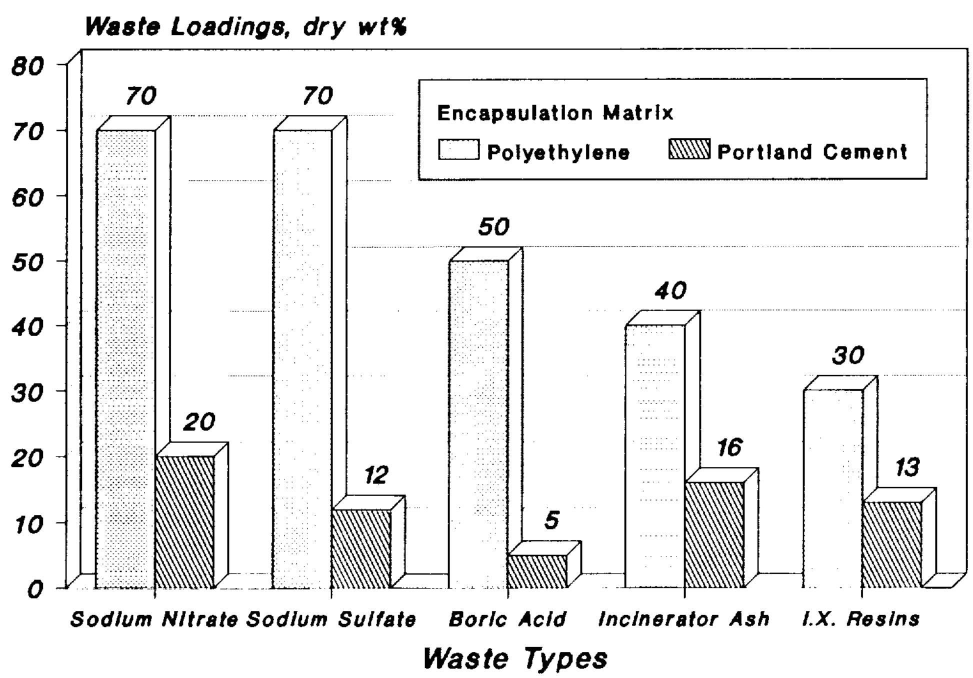


Figure 2 Maximum waste loadings (wt%) for polyethylene encapsulation compared with hydraulic cement processes.

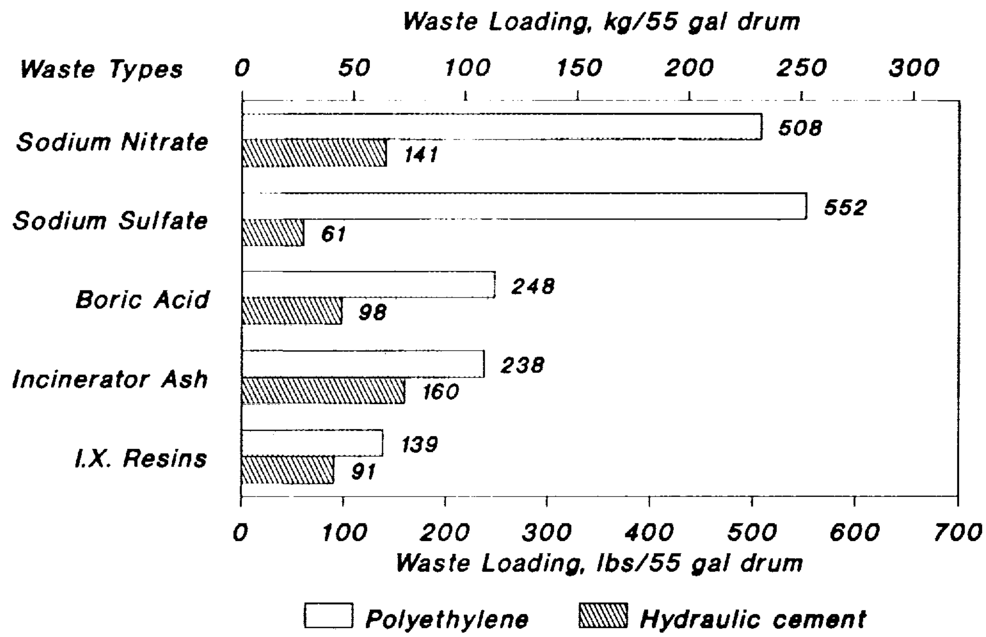


Figure 3 Maximum waste/drum for polyethylene encapsulation compared with hydraulic cement processes.

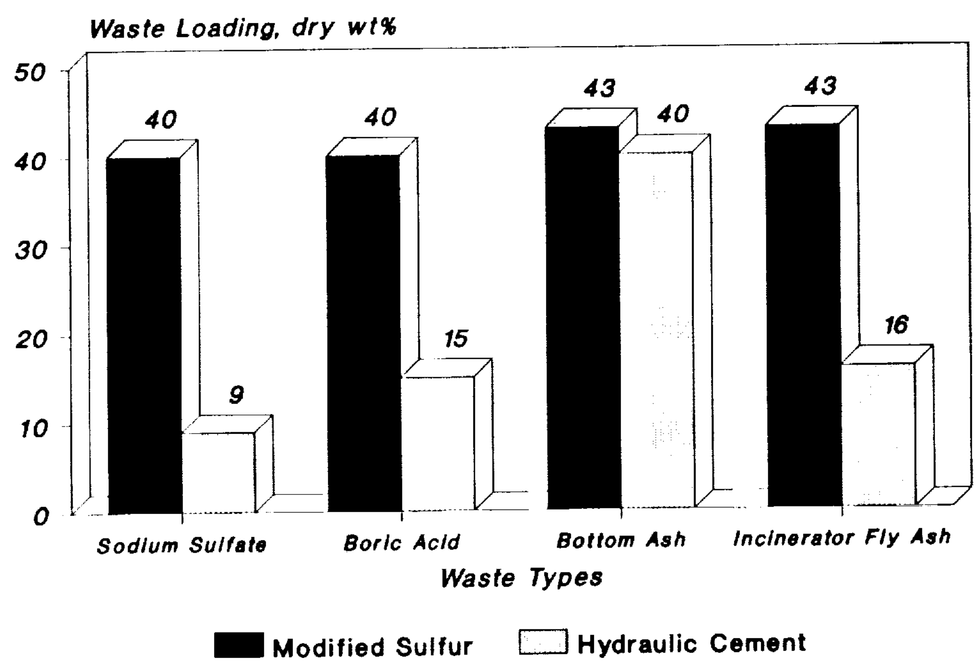


Figure 4 Maximum waste loadings (wt%) for modified sulfur cement compared with hydraulic cement processes.

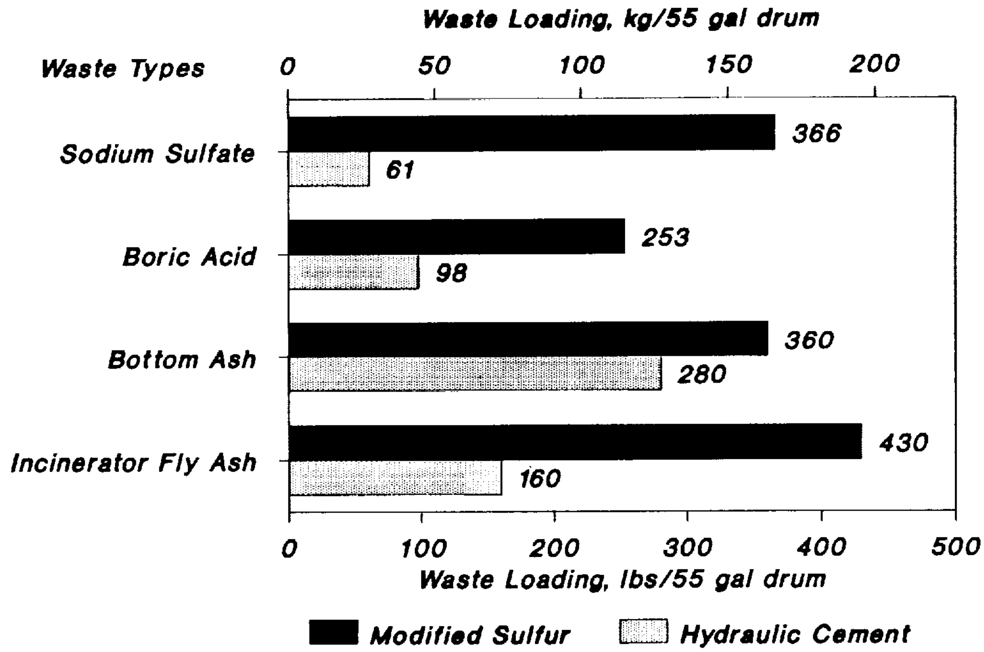


Figure 5 Maximum waste/drum for modified sulfur cement compared with hydraulic cement processes.

PROCESS SCALE-UP AND DEMONSTRATION

On satisfactory completion of bench-scale R & D, scale-up feasibility was investigated for polyethylene encapsulation of nitrate salt wastes. A production-scale feasibility test was conducted using a commercial-grade polyethylene extruder with a 4.5 in. (114 mm) diameter screw and an output capacity of about 2,000 lbs/hr (900 kg/hr). Technical grade sodium nitrate was used to simulate actual mixed waste nitrate salts at a waste loading of 60 wt%. Figure 6 is a process flow diagram for the production-scale test indicating typical parameters. A 30 gallon drum (114 liter) of encapsulated "waste" was filled in about 25 minutes for an average flow rate of about 72 gal/hr (273 l/hr). The resulting product was sectioned to inspect for potential void formation, verify homogenous mixing and provide test specimens for additional confirmatory performance testing. Results of the feasibility test and performance testing are presented in Reference [3], but can be summarized in the following points:

- Polyethylene encapsulation of at least 60 wt% nitrate salt wastes can successfully be accomplished using a production-scale extruder,
- Bench- and production-scale process data are in close agreement,
- QA/performance testing of the 30 gal. waste form demonstrates that a homogenous product with excellent performance properties can be produced using off-the-shelf production equipment.

Based on these results, a production-scale extruder was procured for a technology demonstration to be held at BNL during this fiscal year. The demonstration will be conducted using either actual mixed waste from a DOE facility or surrogate waste that closely approximates actual waste in both chemical and physical composition. This demonstration will be open to all interested parties including those from DOE, NRC, EPA, and the commercial sector.

Scale-up activities for the modified sulfur cement process are continuing and demonstration of production-scale feasibility is planned by the end of FY-1992. The focus of this demonstration will be treatment of mixed waste incinerator fly ash generated at both DOE and commercial facilities.

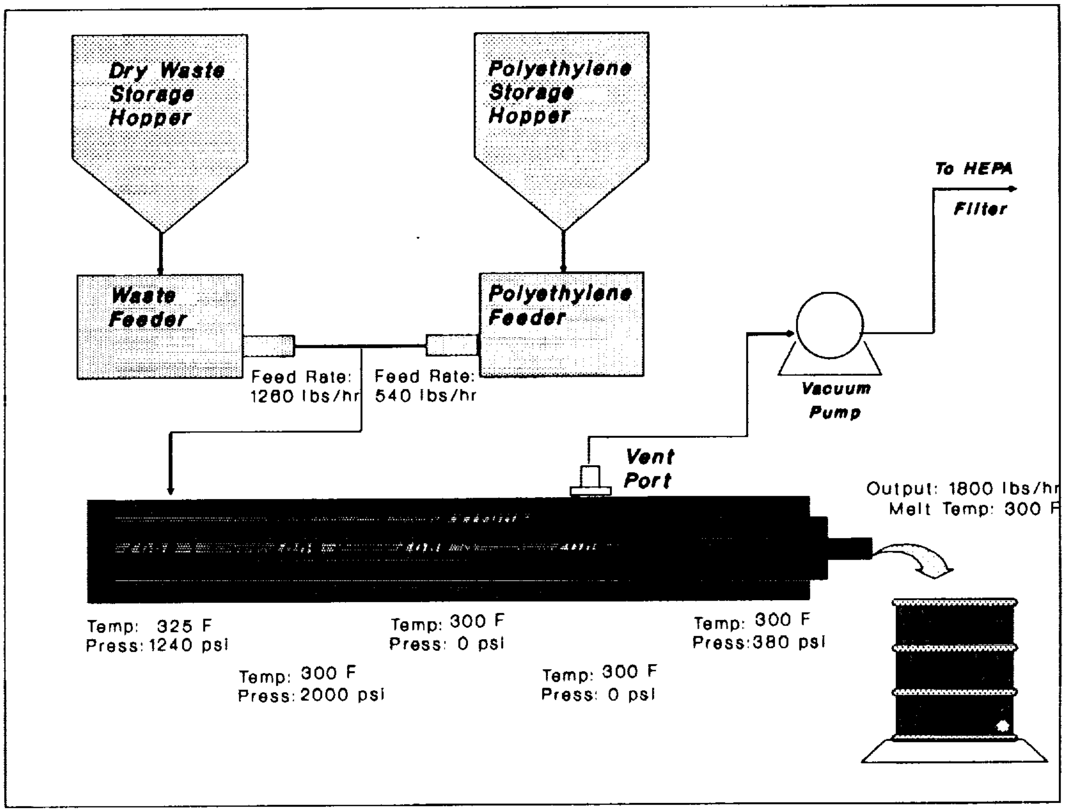


Figure 6 Process flow diagram for full-scale polyethylene encapsulation system.

SUMMARY AND CONCLUSIONS

Waste management technology development and demonstration efforts conducted at BNL are aimed at providing improved methods of treating radioactive, hazardous, and mixed wastes. In keeping with DOE OTD policy, processes must be "better, faster, cheaper, and safer" than conventional technologies. Two systems developed to date (polyethylene and modified sulfur cement encapsulation) have been shown to provide better waste form performance under long-term disposal conditions and improved waste loadings, on a cost-effective basis.

REFERENCES

1. U.S. Department of Energy, "Integrated Data Base for 1988: Spent Fuel and Radioactive Waste Inventories, Projections, and Characteristics," DOE/RW-0006, Rev. 4, Oak Ridge National Laboratory, Oak Ridge, TN, September 1988.

2. Kalb, P.D., J.H. Heiser, and P. Colombo, "Comparison of Modified Sulfur Cement and Hydraulic Cement for Encapsulation of Radioactive and Mixed Wastes," Proceedings of the Twelfth Annual U.S. DOE Low-Level Waste Management Conference, CONF-9008119-Proc., Chicago, IL, August 28-29, 1990.

3. Kalb, P.D., J.H. Heiser, and P. Colombo, "Polyethylene Encapsulation of Nitrate Salt Waste: Waste Form Stability, Process Scale-Up, and Economics," BNL-52293, Brookhaven National Laboratory, Upton, NY, July 1991.

4. McBee, W.C., T.A. Sullivan, and B.W. Jong, "Development and Testing of Superior Sulfur Concretes," RI-8160, Bureau of Mines, U.S. Dept. of Interior, Washington, DC, 1976.

5. Kalb, P.D., J.H. Heiser, and P. Colombo, "Durability of Incinerator Ash Waste Encapsulated in Modified Sulfur Cement," Thermal Treatment of Radioactive, Hazardous Chemical, Mixed and Medical Wastes, Proceedings of the 1991 Incineration Conference, Knoxville, TN, May 13-15, 1991.

6. Kalb, P.D., and P. Colombo, "Polyethylene Solidification of Low-Level Wastes, Topical Report," BNL-51867, Brookhaven National Laboratory, Upton, NY, October 1984.

7. Franz, E.M., J.H. Heiser, and P. Colombo, "Solidification of Problem Wastes, Annual Progress Report," BNL-52078, Brookhaven National Laboratory, Upton, NY, February 1987.

8. Kalb, P.D. and P. Colombo, "Modified Sulfur Cement Solidification of Low-Level Wastes, Topical Report," BNL-51923, Brookhaven National Laboratory, Upton, NY, October 1985.

N92-22457

REGULATED BIOLUMINESCENCE AS A TOOL FOR BIOREMEDIATION PROCESS MONITORING AND CONTROL OF BACTERIAL CULTURES

Robert S. Burlage
Environmental Sciences Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831

Armin Heitzer
Center for Environmental Biotechnology
University of Tennessee
Knoxville, TN 37932

Philip M. DiGrazia
Center for Environmental Biotechnology
University of Tennessee
Knoxville, TN 37932

ABSTRACT

An effective on-line monitoring technique for toxic waste bioremediation using bioluminescent microorganisms has demonstrated great potential for the description and optimization of biological processes. The lux genes of the bacterium Vibrio fischeri are used by this species to produce visible light. The lux genes can be genetically fused to the control region of a catabolic gene, with the result that bioluminescence is produced whenever the catabolic gene is induced. Thus the detection of light from a sample (monoculture, consortium, or bioreactor) indicates that genetic expression from a specific gene is occurring. We have used this technique to monitor biodegradation of specific contaminants from waste sites. For these studies, fusions between the lux genes and the operons for naphthalene (nah) and toluene/xylene (xyl) degradation were constructed. Strains carrying one of these fusions respond sensitively and specifically to target substrates. Bioluminescence from these cultures can be rapidly measured in a non-destructive and non-invasive manner. The potential for this technique in this and other biological systems is discussed.

INTRODUCTION

The development of bioreporter genes is one of the reasons for the rapid advances in the fields of genetics and physiology. A bioreporter gene is associated with a convenient assay, and can substitute for another gene that can not be assayed. The more useful bioreporters are associated with a biochemical assay that has any or all of the following characteristics: the test should be rapid and uncomplicated, inexpensive, reliable, sensitive, and should be free of any background signal. The best bioreporters will incorporate many of these characteristics, although all are lacking in some important facet.

The bioreporter that is used most frequently is the lacZ gene. This gene encodes the enzyme beta-galactosidase, which converts the sugar lactose into glucose and galactose. Its utility as a bioreporter arises from the use of synthetic substrates which produce easily measured colored products after enzymatic action of beta-galactosidase (5). These simple assays have been used in countless experiments to describe genetic expression; however, there are important limitations. Some of the reagents are expensive and the test, although fairly rapid, requires the destruction of part of the sample. Bioreporters which overcome these limitations are often needed.

A bioreporter that would avoid these drawbacks could rely on the measurement of a physical parameter, such as light. Light is produced in a biological reaction called bioluminescence by the marine microorganism, Vibrio fischeri, which lives symbiotically with certain species of deepsea fish. The intricacies of this symbiotic relationship are not well understood, but the genes for this trait (lux) have been successfully cloned (2). Extensive genetic analysis of the lux genes has resulted in the use of a subclone (luxCDABE) to produce light in other bacterial species. The luxAB genes encode a heterodimeric luciferase enzyme that converts an aldehyde group to a carboxyl group on a suitable substrate. This reaction generates visible light as a byproduct. The luxCDE genes convert the carboxyl group back to the aldehyde, so that a pool of substrate is available for continuous light production. The former reaction is dependent on the presence of molecular oxygen, and so only aerobic reactions can be monitored using the lux genes as a bioreporter (4).

In a typical genetic construction the lux genes are fused to the control region (promoter) of the target gene. All of the regulatory genes necessary for the correct functioning of the target gene must also be supplied. The induction conditions for the gene under study will also induce the lux genes, and bioluminescence will result. If conditions are present that are known to cause induction, but no light production occurs, it may be assumed that some inhibitory condition or compound is also present in the reaction milieu. It is thus possible to use a lux fusion to study optimal expression conditions of a particular gene as well as to find appropriate conditions of expression in environmental samples. Often these fusions are located on plasmids - small, circular DNA sequences - that are easily introduced into bacterial species.

The efficacy of this technique in describing appropriate expression conditions, and the use of those conditions with environmental samples, is demonstrated here in a study of the degradation of hazardous waste compounds. The genes responsible for naphthalene catabolism (nah) and for toluene/xylene catabolism (xyl) from the soil bacterium Pseudomonas putida were fused to the lux genes to create sensitive and specific bioreporters. The success of this technique in defining induction conditions is evident in the correlation of waste degradation with periods of maximum light production. This technology should be more generally applicable to other aerobic bacterial processes. It is expected that increased efficiency of bioremediation as well as other processes will result from an analysis of bioluminescent reporter induction.

MATERIALS AND METHODS

Bacterial strains and plasmids. All genetic manipulations were performed in an Escherichia coli DH5 strain. Plasmid constructions were introduced into Pseudomonas putida PB2440 by conjugation. The plasmid pUTK9 (1) contains a nah-lux fusion that was constructed from the promoter region of the NAH7 naphthalene catabolic plasmid and the lux cloning vector pUCD615. The plasmid pUTK24 (Burlage et al., manuscript in preparation) contains a xyl-lux fusion constructed using the promoter region from the TOL toluene/xylene catabolic plasmid and pUCD615 (Figure 1).

Light detection. Light was detected and quantified using an Oriel photomultiplier model 77761 with a liquid light pipe and a collimating beam probe. This apparatus reports light as amperes of induced current, and is usually reported as nanoamps. The photomultiplier probe was kept in a light-tight chamber to reduce incident light.

Media and reagents. All reagents were greater than 99% pure. Naphthalene was added in a crystalline form. Toluene was added as a liquid at a final concentration of 0.015 mM. Relevant incubation conditions for experiments are given in the figure legends. Media in these experiments and protocols for genetic manipulations have been described (7).

RESULTS AND DISCUSSION

Genetic constructions. The essential characteristic of all the bioreporters described here is that they are the product of recombinant DNA manipulations. This is possible because the lux genes have been isolated on a convenient plasmid cloning vector called pUCD615 (6). This plasmid contains the luxCDABE structural genes, but does not carry a promoter (control) region that is essential for the expression of lux. Therefore a promoter region must be introduced into an appropriate position near the lux genes. All the techniques involved in this work are standard in the field of molecular biology, and are fairly inexpensive to perform or to purchase.

An example of one of the bioreporter constructions is presented in Figure 1. This is the cloning scheme for the production of plasmid pUTK24, a bioreporter of toluene and xylene presence. The TOL plasmid actually has the genes for the degradation of these compounds, and the regulatory genes xylS and xylR. We utilized a small fragment of this large plasmid that contains the promoter of the catabolic genes. The new construction, pUTK24, was initially propagated in an E. coli strain, and then moved to a Pseudomonas strain. The new strain, RB1401, also contains the intact TOL plasmid, which provides the two regulatory genes and which allows the strain to degrade the contaminants. Thus this strain can both degrade toluene and report on its presence simultaneously.

It is entirely possible to make a vast number of these constructions, testing for myriad substances or conditions. Generally speaking, it is only necessary to obtain the promoter region of interest (often these can be isolated experimentally), the appropriate host regulatory genes (xylS and xylR in the above example) and the inducer substrate (toluene or xylene above). Other examples will be mentioned in the Discussion section, as well as possible uses.

Induction of bioluminescence. Figure 2 illustrates the generation of visible light by one of these bioreporters after induction with a specific substrate (1). The Pseudomonas strain RB1351 used here contains the nah-lux fusion plasmid pUTK9 and the intact NAH7 plasmid. The strain is able to both degrade and report on the presence of naphthalene. Light is measured in amperes as outlined in the Methods section. Within only a few minutes after naphthalene crystals were added to the lid of the plate, the light production from RB1351 has increased, and a few minutes later light production has reached a maximum value. It remains at this plateau as long as naphthalene crystals were observed (in this case, more than 16 hours). The response of the bioreporter is rapid to a specific inducer molecule, and can be measured in real time due to the on-line characteristic of the assay.

In liquid cultures the induced RB1351 strain gave a result that was different in significant details (1). Although the naphthalene was added while the culture was growing rapidly, the light production did not occur until the growth rate slowed down. It was later shown that naphthalene degradation was always correlated with light production, and that this catabolic system was under a growth-rate regulation. This was an unexpected finding, and was significant because it uncovered an important facet for the optimization of bioremediation in complex systems.

The sensitivity of these reporter strains has also been examined. This is important because contaminants of soil and groundwater are often present in low yet biologically significant amounts. For the nah-lux system the lower limit of detection is at least 45 part-per-billion (ppb) and has been reported as low as 0.1 ppb.

Bioluminescent bioreporters for environmental samples. Both nah-lux and xyl-lux plasmids were used in a recent study of the presence of specific contaminants in soil and water samples from sites near a fuel oil storage facility (Heitzer et al., manuscript in preparation). Typical results are shown in Figure 3, which describes data collected using the xyl-lux plasmid. Two representative soil samples are presented, TP01-08 and TP04-65. TP04-65 clearly shows induction of light, while TP01-08 demonstrates no light production. The former sample came from a site that was obviously contaminated with fuel oil, while the latter came from a relatively clean site. These results show that the contaminated site must have toluene or xylene as a fraction of the total contamination, since the reporter is specific for these compounds. It also demonstrates that this contaminant is **bioavailable** to the reporter strain. This is a very important quality for in situ work because the compounds must not only be present in the soil for bioremediation to have an effect, but those compounds must also be taken up by the bacterial cells and induce the appropriate catabolic genes.

It is evident from this experiment that bioreporters may be useful for a variety of aerobic processes besides bioremediation. The bioreactor can contain a consortium or a pure culture, the feedstock can be constant or variable. The bacterial strain that is actually performing the enzymatic activity is also reporting on its progress. In the case cited above, the light output continues until the toluene substrate is exhausted i.e. until it falls below a level sufficient for induction. This information could be very useful in designing a bioreactor system or optimizing the parameters of a particular reaction. The utility of this technique for complex bioreactor was recently described in an examination of bioreporter response to a variable feed stream (3).

Projects in development. Several other lux constructions have been created in this laboratory. A fusion to the genes for mercury reduction (mer) has produced a bioreporter of heavy metal contamination. Other constructions include constitutive light producers and bioreporters of stress conditions. Plans have also been made to mutagenize the lux genes in an attempt to create bioluminescence with different colors (wavelengths). These could be used to report on several genes in the same strain or in the same consortium.

The versatility of lux bioreporter strains makes them ideal for a variety of purposes. Experiments are in progress to optimize expression from the xyl-lux fusion in a bioreactor. The goal of this research is not only to optimize degradation of toluene and xylene, but also to learn about the physiology of the Pseudomonas bacterium and eventually to study the complex microbial interactions in a mixed bacterial culture. The knowledge gained from this series of studies may eventually expand the scope of microbial products and processes.

An unusual assimilation of microbial genetics, light detection apparatus and integrated computer technology will be used to control and modify bioreactor conditions. In this scheme, the bioluminescence from a reactor vessel will be detected by a photomultiplier which is monitored by a computer. If the light production falls below a certain threshold value the conditions are assumed to be no longer optimal, and a computer subprogram will be activated to adjust the process conditions. A series of interfaces with diagnostic equipment will allow the computer to determine whether key operating parameters are within normal ranges. When a non-optimal condition is discovered the conditions are appropriately modified. An appropriate length of time is allowed to pass for recovery of the bioluminescence, and the light is again sampled. If the output is normal the computer resumes normal sampling. If light output is still below threshold the diagnostic subprogram searches for other parameters. Certain key components of this system are already functional. When all units are in place it is believed that an efficient bioprocess control will be achieved.

SUMMARY

Our initial experiments were designed to determine the suitability of lux gene bioreporters for use as indicators of genetic expression under defined conditions. The acceptance of a new bioreporter by the scientific community is difficult, since the new bioreporter must offer advantages that are unavailable with other common systems. The lux system supplies many advantages that make it attractive for broad use, both in academia and in industry.

The specificity of the bioreporter for one or a few inducer substrates is an advantage inherent in using a genetic fusion as an indicator. Many such bioreporter strains can be constructed, not only for bioremediation of hazardous waste, but for any aerobic process. This might include industrial production of valuable recombinant proteins, antibiotics, biomass, or other products. Applications are only limited by worker availability and the description of suitable promoters.

The sensitivity of the lux system is also a great asset. It is possible to accurately measure light at extremely low intensities, and this means that very few cells can be detected. This makes the lux system valuable for bioreactor studies in which the bioluminescent strain may be present in low numbers, yet still be an important component of the consortium. This also means that in situ work might eventually be possible, so that these bioreporters might be portable sensors of environmental conditions.

Other important characteristics of this system include the speed of the assay. The results presented here were obtained on a real-time basis, allowing instant analysis of perturbations to the system. The ease with which the assay was performed allowed a many measurements to be taken cheaply, where other bioreporters would require selection of discrete timepoints as representative of the culture response. In addition, none of the sample was sacrificed for the assay, and the reactor was not disturbed during the procedure.

The experiments in progress should lead to new strategies for designing effective biological processes and for engineering of reactor systems. It is anticipated that further refinements will make the lux system more versatile, more sensitive and powerful, and more popular with the business and academic communities.

ACKNOWLEDGEMENTS

This research was supported by the U.S. Department of Energy's Office of Environmental Restoration and Waste Management, Office of Technology Development. Oak Ridge National Laboratory is managed by Martin Marietta Energy Systems, Inc. under contract DE-AC05-84OR21400 with the U.S. Department of Energy. Publication #XXXX of the Environmental Sciences Division, Oak Ridge National Laboratory.

REFERENCES

1. Burlage, R.S., G.S. Sayler, and F. Larimer. 1990. Monitoring of naphthalene catabolism by bioluminescence with nah-lux transcriptional fusions. J. Bacteriol. 172: 4749-4757.

2. Engebrecht, J., K. Nealson, and and M. Silverman. 1983. Bacterial bioluminescence: isolation and genetic analysis of functions from Vibrio fischeri. Cell 32: 773-781.

3. King, J.M.H., P.M. DiGrazia, B. Applegate, R. Burlage, J. Sanseverino, P. Dunbar, F. Larimer, and G.S. Sayler. 1990. Rapid, sensitive bioluminescent reporter technology for naphtahlene exposure and biodegradation. Science 249: 778-781.

4. Meighen, E.A. 1991. Molecular biology of bacterial bioluminescence. Microbiol. Rev. 55: 123-142.

5. Miller, J.H. 1972. Experiments in Molecular Genetics, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

6. Rogowsky, P.M., T.J. Close, J.A. Chimera, J.J. Shaw, and C.L. Kado. 1987. Regulation of the vir genes of Agrobacterium tumefaciens plasmid pTiC58. J. Bacteriol. 169: 5101-5112.

7. Sambrook, J., E.F. Fritsch, and T. Maniatis. 1982. Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

FIGURE LEGENDS

Figure 1. Construction of a lux fusion plasmid. The lux plasmid vector pUCD615 is opened by the restriction enzyme EcoRI. The promoter (P) region of the TOL plasmid is isolated using EcoRI. The promoter fragment and the cut plasmid are combined and fused together using the ligase enzyme, creating the plasmid pUTK24. When this plasmid is introduced into a Pseudomonas strain, the strain becomes a bioreporter of toluene and xylene bioavailability. xyl -genes for toluene, xylene catabolism; lux - genes for bioluminescence.

Figure 2. Bioluminescence is induced by the presence of naphthalene. A nah-lux bioreporter strain was grown on a plate of LB agar until mature colonies formed. The low constitutive expression of light can be seen between 0 and 0.35 hours. Immediately after 0.35 hours naphthalene crystals were added to the lid of the plate. Light is measured by a photomultiplier and reported in nanoamps as described in the text.

Figure 3. The xyl-lux bioreporter strain is used to detect contaminants in soil samples. One gram soil samples were suspended in 4 ml of minimal medium in a small vial. The bioreporter strain was added at a concentration of 10^8 bacteria per ml. The positive control contained added toluene at a concentration of 0.015 mM; the negative control contained an uncontaminated soil sample. Light is reported in nanoamps.

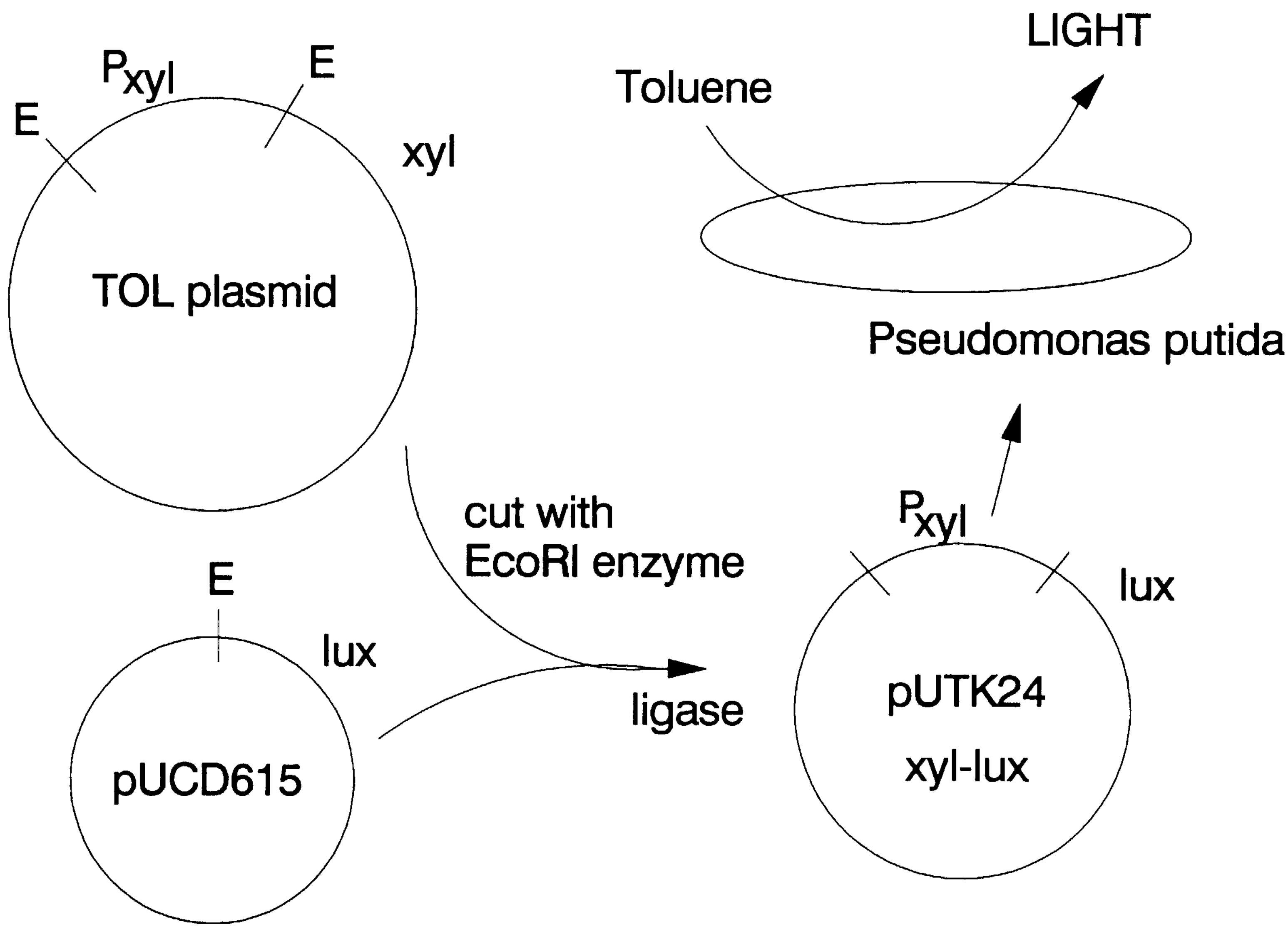
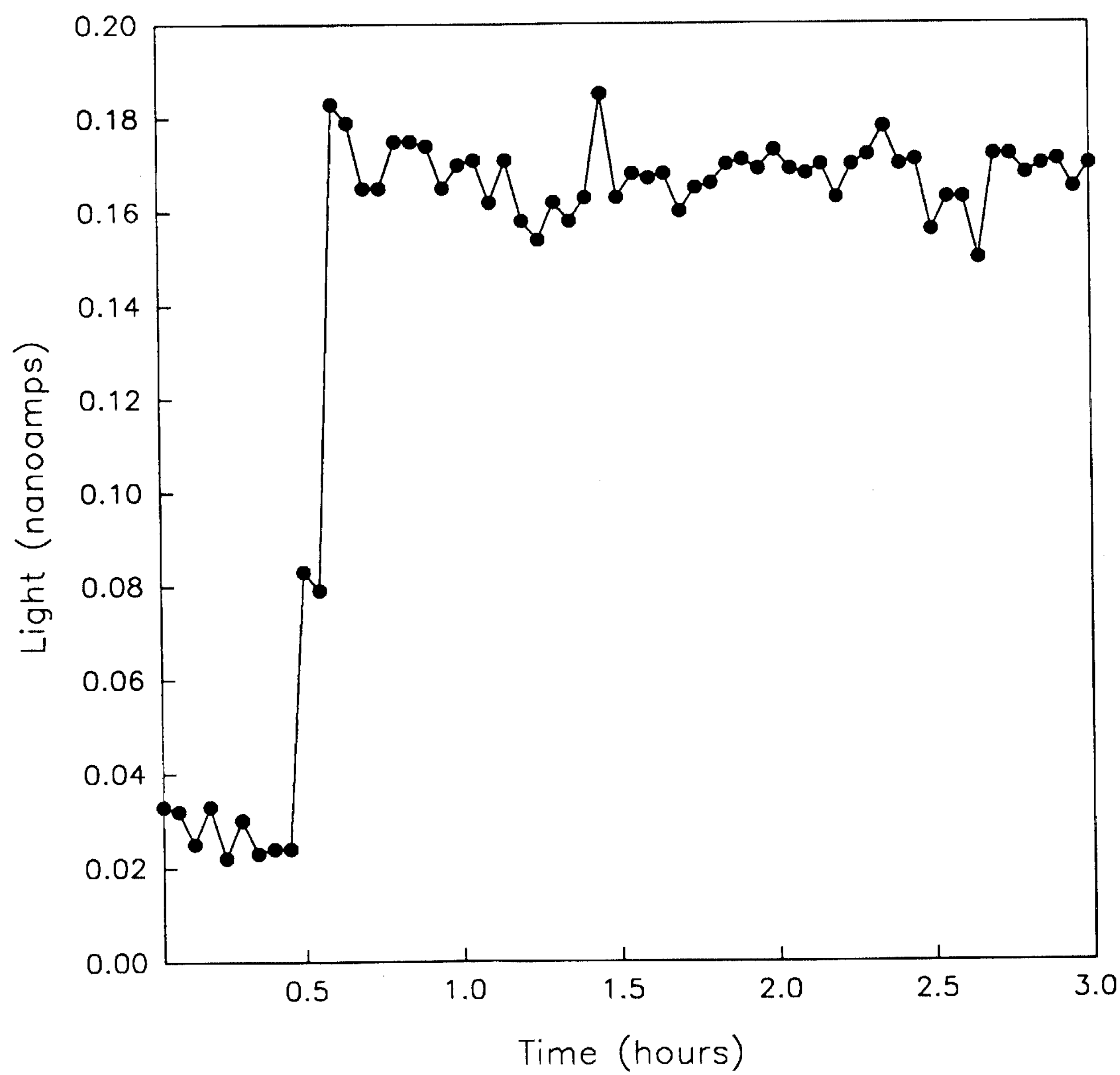
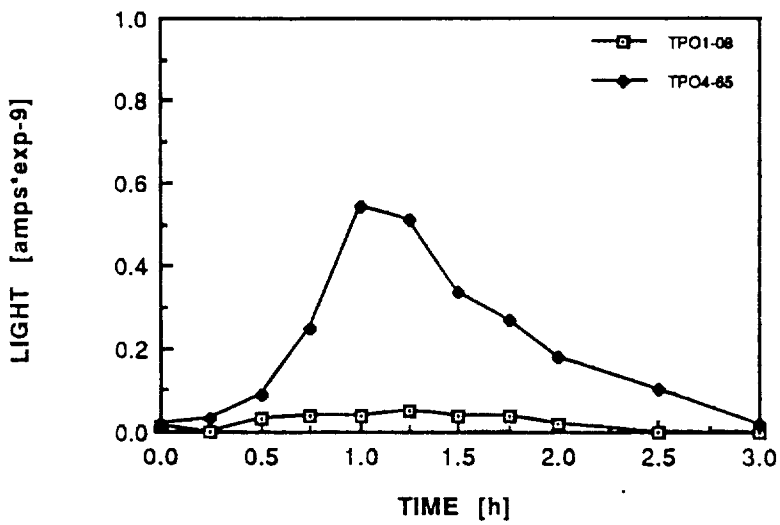


FIGURE 1 - lux plasmid construction

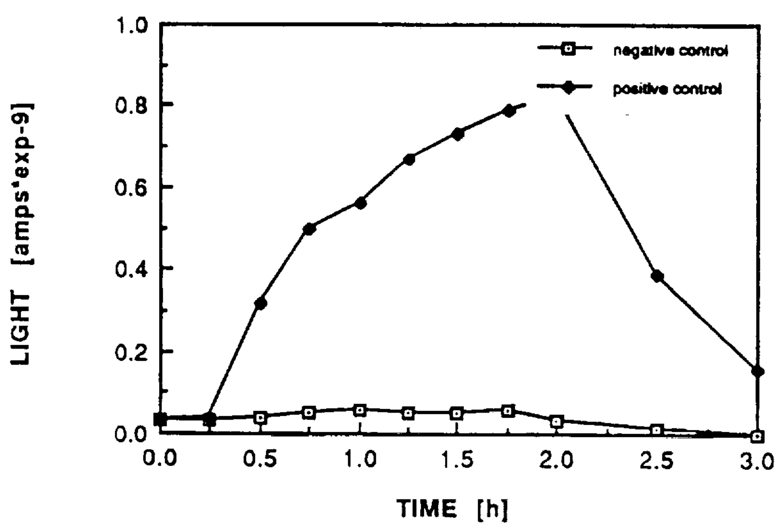
FIGURE 2 – Bioluminescence induced by naphthalene



Expression of the toluene operon in RB1401



RB1401 controls



N92-22458

FIBER OPTIC-BASED BIOSENSOR

Frances S. Ligler
Center for Bio/Molecular Science and Engineering,
Naval Research Laboratory,
Washington, DC 20375-5000

ABSTRACT

The NRL fiber optic biosensor is a device which measures the formation of a fluorescent complex at the surface of an optical fiber. Antibodies and DNA binding proteins provide the mechanism for recognizing an analyze of interest and immobilizing a fluorescent complex on the fiber surface. The fiber-optic biosensor is fast, sensitive, and permits analysis of hazardous materials remote from the instrumentation. The fiber optic biosensor is described in terms of the device configuration, chemistry for protein immobilization, and assay development. A laboratory version of the device is being used for assay development and performance characterization while a portable device is under development. Antibodies coated on the fiber are stable for up to two years of storage prior to use. The fiber optic biosensor has been used to measure concentration of toxins in the parts per billion (ng/ml) range in under a minute. Immunoassays for small molecules and whole bacteria are under development. Assays using DNA probes as the detection element can also be used with the fiber optic sensor. The fiber optic biosensor is currently being developed to detect biological warfare agents, explosives, pathogens, and toxic materials which pollute the environment.

INTRODUCTION

A biosensor is a detection system which exploits the sensitivity and selectivity of a biomolecule for analyze recognition and incorporates that biomolecule into an optoelectronic device for signal transduction(1-3). The NRL fiber optic biosensor uses antibodies and DNA probes to provide the recognition component. Upon binding of analyze, a fluorescence signal is generated at the surface of the optical fiber. The fiber-optic biosensor is described here in terms of the device configuration, chemistry for protein immobilization, and assay development.

HARDWARE

The hardware portion of the fiber optic biosensor has four basic components: the light source, the light coupling optics, signal detection devices, and the optical fiber itself. We have selected all components to minimize the noise that can be confused as signal. Figure 1 shows a schematic of the system.

The laser light source was selected to provide moderate power, stability, coherency, and a narrow excitation wavelength. Fluorescent labels such as the rhodamines are excited in the 500 nm range and emit in the 600 nm range where there is little intrinsic fluorescence in most clinical and environmental samples. Thus a light source producing a 500-550 nm wavelength is required. Our laboratory breadboard device (Figure 2) employs a highly stable, air-cooled argon-ion laser, which emits at 514 nm. We have also successfully used a solid state, frequency-doubled YAG laser which emits at 532 nm and is more suited for a portable device. Power outputs of 5 mW are adequate.

The second basic group of components include the launching and coupling optics: including an excitation filter, focusing lens, and fiber positioner. The excitation filter removes plasma lines from the laser source. A spherical fused silica lens focuses the light onto the fiber. This simple lens has proven easier to use than a microscope objective, but either is feasible as long as a sufficient focal length is maintained to provide room for the chopper.

Discrimination of signal from noise begins even before the excitation light enters the fiber. A chopper positioned between the focusing lens and the fiber modulates not only the excitation light, but also fluorescence generated by the optical components in the path of the excitation beam(4). The chopper and photodiode responsible for signal collection are interfaced to the lock-in amplifier to convert the optical signal to an electrical output and decouple the apparatus from ambient light sources. The lock-in subtracts the modulated signal caused by excitation light and background fluorescence from the optics from the total fluorescence signal. This technique significantly improves the signal-to-noise ratio.

The additional devices which separate the fluorescence emitted from the fiber from the excitation light include an off axis-parabolic mirror(4) and a KV550 long-pass filter. The light returning the fiber is defused across the face of the parabolic mirror. Much of the excitation light reflecting off the proximal end of the fiber returns through the hole in the mirror. The mirror refocuses the emitted light, sending it through the emission filter to remove any stray excitation light, and on to the photodiode. A photodiode was selected, rather than a photomultiplier tube, because of low cost, reliability and compatibility with the lock-in amplifier.

The last, but certainly not the least, of the four basic components of the fiber optic biosensor is the optical fiber. In order to perform homogeneous assays, the fiber is configured to measure signal generated in the evanescent wave (Figure 3) (5). The cladding is stripped from 5-10 cm of fiber at the distal end and the end of the fiber is blocked with black epoxy. Light actually travels out of the fiber into the surrounding aqueous media approximately 100 nm. Fluorescence in preferentially coupled into this evanescent wave and reenters the fiber. Light does not reach the fluorophores in the bulk solution and thus they generate no fluorescence. This geographic discrimination enables separation of signal from sample-derived noise without a physical separation procedure (i.e. washing or filtration).

The fibers we use are cut at a length of approximately 1 meter for convenient handling lengths up to 20 meters have been tested. The proximal end of the fiber is glued into a connector for easy insertion into the sensor. The fiber end is polished to a flat, smooth surface for efficient launching of light into the fiber. The polishing step makes the face of the metal connector shiny. To decrease noise for the shiny surface, the connector face is roughened with acid and blackened with flat enamel paint.

The distal or "business" end of the fiber receives a more intensive preparation. The fibers that are decladded in the last 5-10 cm, some of the emission light traveling back up the fiber is lost at the interface between the sample solution and the cladding. To prevent this signal loss, a method for propagating the signal nearer to the center of the fiber was developed(2). Several fiber geometries have tested, but the one that yields the most power in the evanescent wave, the least excitation of bulk fluorescence, and the best propagation of emitted light is a continuous taper. A computer-controlled immersion in hydrofluoric acid is used to continuously decrease the diameter of the unclad portion of the fiber from 200 microns to 100 microns.

IMMOBILIZATION CHEMISTRY

Once the fiber is tapered, the proteins responsible for detection are immobilized on the unclad surface of the distal end. As described elsewhere(5,6), the fibers are cleaned and coated with a thiol-silane. A heterobifunctional crosslinker, which reacts at one end with thiol groups and at the other end with terminal amino groups, is used to attach antibodies or DNA-binding proteins covalently to the fiber surface (Figure 4). Antibodies are routinely immobilized at 2 ng/mm^2 (2) with a capacity to bind large protein antigens at a ratio of 1 antigen per 2 antibodies. Antibodies immobilized using the procedure maintain antigen binding capability for up to 19 months.

ASSAYS

Three types of immunoassays and one DNA probe assay are being developed for use with the fiber optic-based biosensor. The immunoassays include those specific for small molecules, proteins, and bacteria. To develop an assay for small molecule, antibodies to trinitrophenol are immobilized and a competitive assay to detect trinitrotoluene (TNT) developed. For detection of larger molecules such as proteins, antibodies to botulism toxin are immobilized on the fiber and a sandwich immunoassay developed with fluorescent antibodies in solution. Sensitivities of a ng/ml are obtained. To detect bacteria, ant-salmonella antibodies are immobilized and the probe exposed to a preparation of nonspecifically stained cells. The cells bound by the antibody to the fiber generate a signal. In all of these assays, signal generation begins immediately upon introduction of the probe into the solution containing the fluorescent reagent. Fibers coated with irrelevant antibodies are used to control for false positive signals.

A DNA-probe assay has been developed by using the fiber optic biosensor in conjunction with the polymerase chain reaction (PCR)(8). DNA extracted from samples to be analyzed is subjected to PCR in the presence of primers which bind to the DNA sequence unique to the organism or interest. Following sufficient cycles for million-fold amplification of the selected DNA (20-70 min), the DNA is subjected to a second, shorter PCR using nested primers. In this next replication, fluorescent nucleotides and a sequence identifiable by a DNA-binding protein are incorporated into the amplified DNA. The amplified DNA is then introduced to a fiber optic probe coated with the DNA binding protein. The fluorescence signal is generated in minutes as compared to the hour that would be required for electrophoretic identification of the PCR product.

CONCLUSION

NRL has developed a novel fiber optic biosensor which can use long fibers for analysis remote from the optical components. The optics minimize optical noise and are amenable to miniaturization. The fiber-optic probe is configured for homogeneous assays using the evanescent wave and its geometry significantly improves the signal transmission back up the fiber. Chemistry for immobilizing proteins on the fiber has been developed that successfully immobilize a high density of functional molecules. Antibodies can be immobilized on the fiber up to two years before use. A wide variety of assays can be developed with transduce a binding event into a fluorescent signal. Small and large molecules and pathogenic organisms have been detected in less than a minute.

NOTES AND ACKNOWLEDGEMENTS

Four patents have been filed covering this technology and are available for license(4,5,7,8). The work would not have been possible without support from the Office of Naval Technology and the U.S. Marine Corps. The authors particularly thank Lisa Shriver-Lake, Robert Ogert, Richard Thompson, Carl Villaruel, James Campbell and Steve Walz for their critical scientific contributions. The views expressed here are the authors own and do not reflect policy of the U.S. Navy, Department of Defense or United States Government.

REFERENCES

1. Buck, R.P. et.al., eds; Biosensor Technology Marcel Dekker, Inc. NY, 1990, 419 pp.

2. D.L. Wise, L.B. Wingard, Jr., eds; Fiber Optic Biosensors Humana Press, NJ, 1991, 370 pp.

3. Wolfbeis, O.S., ed.; Fiber Optic Chemical Sensors and Biosensors, CRC Press, Inc. Boca Raton, FL, 1991, Vol. I, 385 pp, Vol 2, 341 pp.

4. Thompson, R., M. Levine; Improved fiber optic fluorescence sensor. U.S. Patent application # 07/531721 (Filed 6/1/90).

5. Thompson, R. and C. Villaruel; Waveguide-binding sensor for use with assays. Patent Application # 07/6610895 (Allowed 5/20/91)

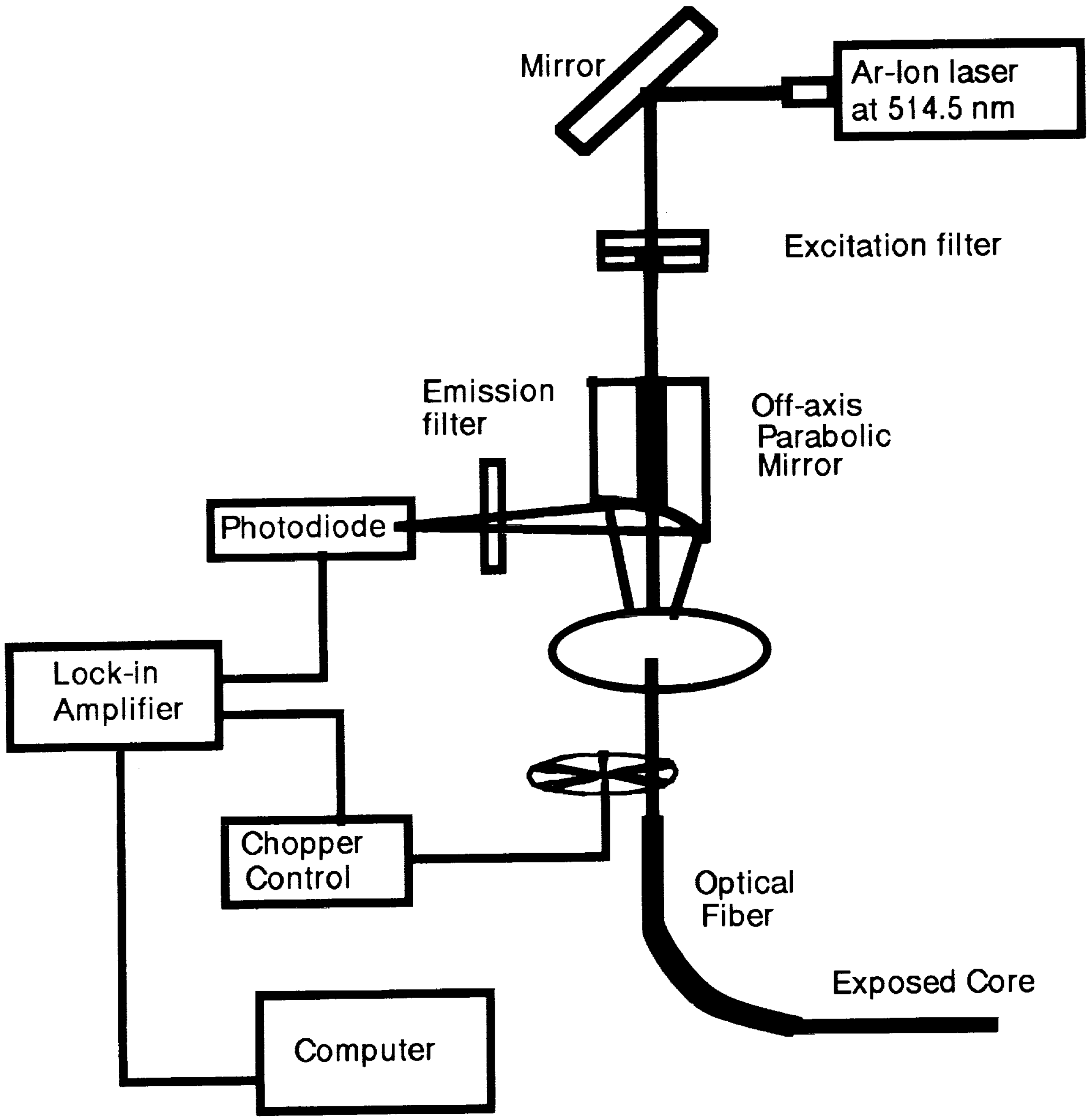
6. Bhatia, S.K., L.C. Shriver-Lake, K.J. Prior, et al.; Use of thiol-terminal silanes and heterobifunctional crosslinkers for immobilization of antibodies on silica surfaces. Anal. Biochem, 178, 408-413, 1989.

7. F.S. Ligler, J. Calvert, J. Georger, L. Shriver-Lake, S. Bhatia, and R. Bredehorst: Means and method for immobilizing active agents on substrates. Patent Application #07/297,088 (Allowed 5/21/91).

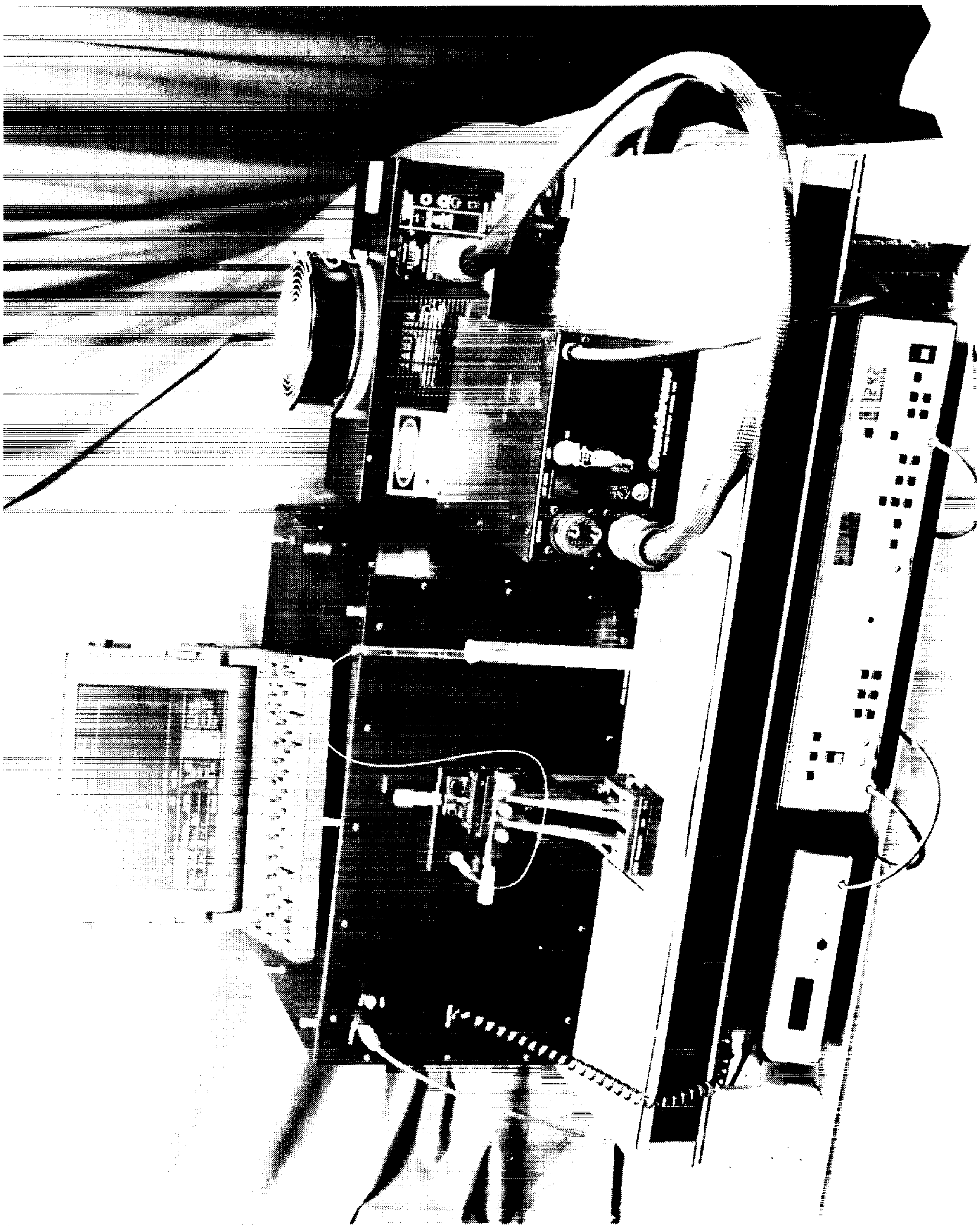
8. Campbell, J.; DNA-based fiber optic sensor. Patent Application #07/635019 (Filed 12/28/90).

FIGURE TITLES

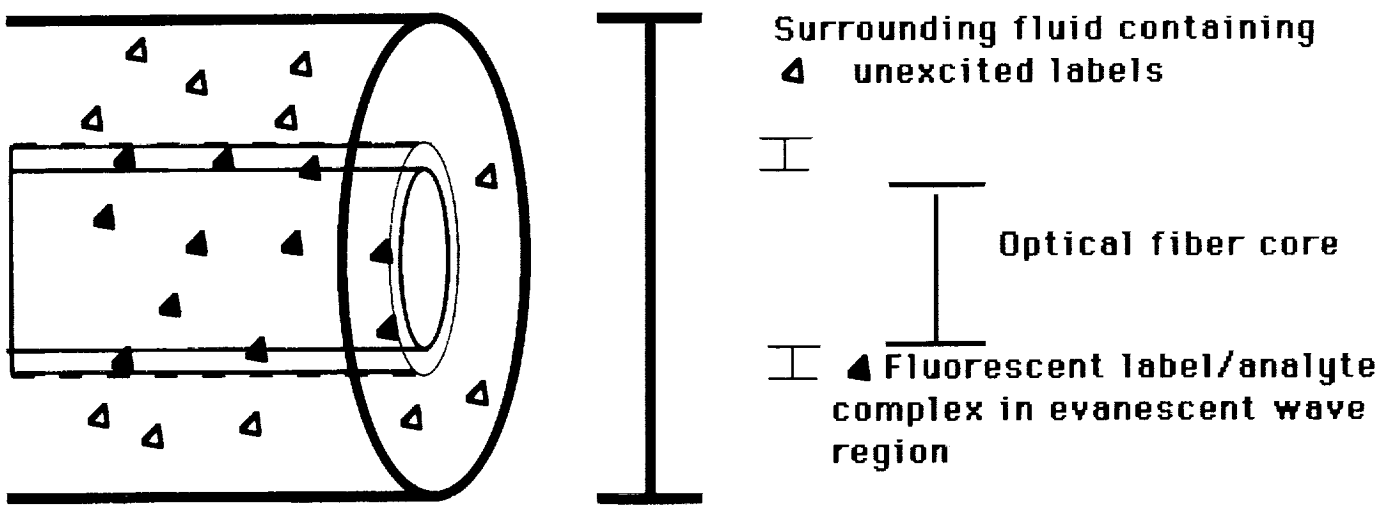
SCHEMATIC DIAGRAM OF THE LABORATORY BIOSENSOR

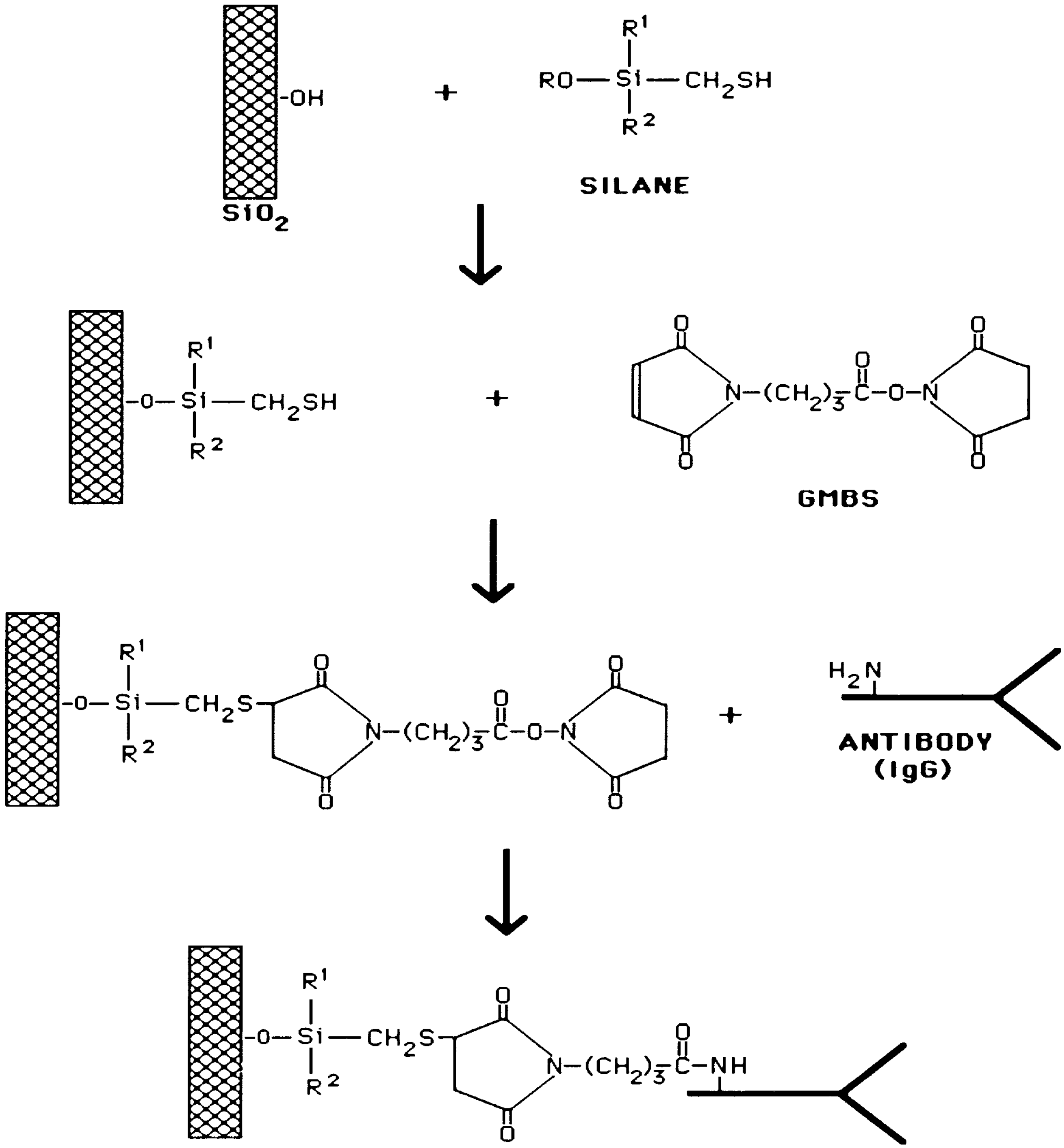


ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH



FIBER OPTIC PROBE IN CROSS-SECTION




RO—Si—CH2SH
R1
R2
-OH
SiO2
SILANE
-O—Si—CH2SH
GMBS
N-(CH2)3-C-O-N
ANTIBODY
(IgG)
H2N
N-(CH2)3-C-NH

N92-22459

AMBIENT-TEMPERATURE CO-OXIDATION CATALYSTS

Billy T. Upchurch
NASA Langley Research Center
Hampton, VA 23665

David R. Schryer
NASA Langley Research Center
Hampton, VA 23665

Kenneth G. Brown
Old Dominion University
Norfolk, VA 23529

Erik J. Kielin
Old Dominion University
Norfolk, VA 23529

ABSTRACT

Oxidation catalysts which operate at ambient temperature have been developed for the recombination of carbon monoxide (CO) and oxygen (O_2) dissociation products which are formed during carbon dioxide (CO_2) laser operation. Recombination of these products to regenerate CO_2 allows continuous operation of CO_2 lasers in a closed-cycle mode. Development of these catalyst materials provides enabling technology for the operation of such lasers from space platforms or in ground-based facilities without constant gas consumption required for continuous open-cycle operation. Such catalysts also have other applications in various areas outside the laser community for removal of CO from other closed environments such as indoor air and as an ambient-temperature catalytic converter for control of auto emissions.

INTRODUCTION

The electrical discharge used to excite many pulsed CO_2 lasers generally causes some decomposition of CO_2 to CO and O_2. This decomposition can be detrimental to long-life laser operation due to both the loss of CO_2 and the buildup of O_2. CO_2 loss causes a corresponding gradual decrease in laser power, but the buildup of relatively small amounts of O_2 can cause rapid power loss and even complete failure in many lasers. In some commercial applications these problems are overcome by operating the lasers open-cycle with a continuous flow-through of fresh laser gas to effect removal of dissociation products. Such open-cycle operation is impractical for space-based applications or other applications involving weight or volume constraints. Consequently, closed-cycle operation with recycling of the laser gas is necessary.

Closed-cycle operation of CO_2 lasers requires a suitable CO-oxidation catalyst to recombine the dissociation products, CO and O_2, to regenerate CO_2. For effective closed-cycle laser operation, the catalyst used must have a high recombination efficiency at low concentrations of O_2 and ambient laser gas temperatures, typically between 25 °C and 75 °C, so additional energy is not required to heat the catalyst. Additionally, applications with rare isotope CO_2 for enhanced atmospheric transmission require utilization of a catalyst whose operation is compatible with the particular isotope chosen without contributing normal isotope impurities. The NASA Laser Atmospheric Wind Sounder (LAWS) will use such a laser requiring five years of continuous operation to measure global winds from orbit. For such long-term operation, this application requires a closed-cycle, pulsed $C^{18}O_2$ laser and a $C^{18}O$-oxidation catalyst which will operate under ambient temperature laser gas conditions without isotopic scrambling. Such CO-oxidation catalysts with high activity at ambient laser gas temperatures and low oxygen concentrations were nonexistent until fairly recently [2-4,11-23].

CO-oxidation catalysts which have been effective at low oxygen concentrations generally have been noble metals such as platinum (Pt) or palladium (Pd) on a support and have shown activity only at elevated temperatures. CO-oxidation catalysts with useful activity at ambient temperature generally require air or other gases which supply high concentrations of O_2 to be effective. The commercial catalyst Hopcalite which is a mixture of nonprecious metal oxides is such a catalyst.

Recently a new class of catalysts has been developed which consists of a noble metal on an active reducible-oxide support. Several of these noble-metal/reducible-oxide (NMRO) catalysts such as platinized tin-oxide (Pt/SnO_2) and gold/manganese dioxide (Au/MnO_2) have been shown to have high activity for CO oxidation at room temperature even with low and stoichiometric concentrations of CO and O_2. Consequently these catalysts are suitable for O_2 and CO removal through CO_2 regeneration in closed-cycle CO_2 lasers. Since the activity of these catalysts increases with increasing O_2 concentration they have very high ambient-temperature activity for CO oxidation in air and are thus potentially useful for air purification. They may also be useful as a catalyst component for internal combustion engine exhaust gas conversion as high gas temperatures are not required for effective operation of the catalyst.

NMRO catalysts, particularly those based on Pt/SnO_2, have been extensively studied at the NASA Langley Research Center (LaRC) over the past several years. While the principle impetus for this research has been CO_2 laser applications related to NASA's LAWS laser instrument development program, research involving application of NMRO catalysts in air purification has been initiated as well. Results of these efforts will be presented and discussed.

EXPERIMENTAL

The Au/MnO_2 and Pt/SnO_2 catalysts were tested in plug-flow or recycle reactors, the details of which have been reported earlier [1-3]. Briefly, the plug flow reactor is designed to simulate laser dissociation conditions by flowing a test gas containing low partial pressures of CO and O_2 in an inert matrix gas over a catalyst sample in a temperature controlled quartz tube and then measuring the resulting conversion to CO_2. Quantitative analysis of the gas composition is accomplished either gas chromatographically or mass spectrometrically in the case of isotopic studies.

Unless otherwise indicated, the catalyst sample sizes were 0.050 grams for monolith supported samples. This value is exclusive of support. The test gases used were one or more of the following: (1) stoichiometric 1.0% CO, 0.50% O_2; (2) stoichiometric with 16% CO_2; (3) twice the stoichiometric amount of CO: 2.0% CO, 0.5% O_2; (4) twice the stoichiometric amount of O_2: 1.0% CO, 1.0% O_2;and (5) 500 ppm CO in air for air purification studies. All test gases except the air purification test gas mixture were in a high purity He matrix with 2 percent Ne added as an internal calibration standard. Conversion efficiencies were measured at 35°C and 55°C unless otherwise specified.

The Au/MnO_2 catalysts in powder form were prepared by Gar Hoflund at the University of Florida using standard coprecipitation techniques. All Pt/SnO_2 samples and Au/MnO_2-coated monolith samples were made at NASA LaRC. Syntheses of Pt/SnO_2 catalysts has been reported earlier [4]. The Hopcalite was 8 mesh commercially available material. Monolith coating techniques are proprietary at present.

DISCUSSION

The first catalyst to show significant near ambient temperature CO-oxidation activity was platinized tin-oxide (Pt/SnO_2), a noble-metal/reducible-oxide (NMRO) catalyst. Platinized tin oxide has been extensively researched to optimize formulation, synthesis, and pretreatment conditions resulting in a significantly improved catalyst [1-12]. Unfortunately, this catalyst loses about half of its initial activity within a few days and then exhibits a slower decay with a half-life of several months. Research into the cause of decay indicates that the initial decay is due to CO_2 retention, possibly forming bicarbonate or carbonate, whereas the long-term decay appears to be associated with changes in surface morphology.

Since the discovery of the synergistic effect of combining Pt and SnO_2 for the enhanced catalytic oxidation of CO by O_2 at near-ambient temperatures, several improvements have been made, not only in Pt/SnO_2 but also in other noble-metal/reducible-oxide (NMRO) catalysts. The research efforts at NASA Langley Research Center in the development of CO-oxidation catalysts have been reported extensively in other articles [1-13]. Improvements over commercially available catalysts resulting from these research efforts include: (1) development of an inherently clean method of preparing catalysts, eliminating chloride contaminants and allowing formerly unprecedented loadings of up to 46% Pt/SnO_2; (2) determination of the optimum ratios of Pt/SnO_2 and $Pd/Pt/SnO_2$ for best activity (17% Pt and 5% Pd by weight); (3) development of a method of coating Pt/SnO_2 onto a high-surface-area, silica gel support providing enhanced activity with minimal humidification of the laser gas; and (4) determination of the optimum reductive pretreatment conditions of time (1.0 hour), temperature (125 °C), and gas (5% CO in He) for best catalyst activity [11].

Other noteworthy discoveries reported earlier are that (1) appropriate pretreatment temperature, moisture, and Pt loading eliminate the initial dip in catalytic activity; (2) bicarbonate and/or carbonate build up contribute to catalyst decay [11,13]; (3) isotopic scrambling of common-isotope ^{16}O from the catalyst surface in the $C^{18}O$-oxidation process can be eliminated by replacing reactive surface oxygen in Pt/SnO_2 with ^{18}O [4,11]; (4) the reaction between O_2 and CO is first order in O_2 and apparent first order for the overall reaction [2,11]. Extensive surface studies of Pt/SnO_2 indicate that a Pt/Sn alloy having surface hydroxyls forms when Pt/SnO_2 is reductively pretreated. Both the alloy and the surface hydroxyls are believed to contribute to the catalytic activity. [14-23].

Recent advances made in CO_2 laser catalysts to be discussed herein include comparisons of the activity of Au/MnO_2 to Pt/SnO_2 catalysts with possible explanations for observed differences. The catalysts were compared for the effect of test gas composition, pretreatment temperature, long term activity, effects of added promoter compounds, and surface labeling for isotopic laser gas compatibility, and applicability for use as ambient temperature air purification catalysts.

As shown in Figure 1 the activity of the Au/MnO_2 catalyst decreases significantly when the O_2 is less than stoichiometric and increases dramatically under O_2-rich conditions. Pt/SnO_2 behaved similarly. The activity enhancement from O_2 is expected as the reaction is first order in O_2. The catalysts do not exhibit a simple reaction order with respect to CO. The apparent CO order is zero at high excesses of O_2 and at stoichiometric levels. It is probable that CO adsorption competes with O_2 for active sites thus giving the observed reduced activity [13]. Thus, in air these catalysts could function in air purification applications.

In the absence of CO_2, Au/MnO_2 possesses both higher activity and lower decay than Pt/SnO_2. As shown in Figure 2, however, when CO_2 is present at high concentrations in the test gas, the activity of the Pt/SnO_2 remains relatively unaffected while that of the Au/MnO_2 falls below that of the Pt/SnO_2. These results indicate that while the Au/MnO2 would not be a viable candidate for CO_2 laser applications, both catalysts could function in gas purification applications in which the CO_2 concentration is low.

A promoter compound has recently been found which significantly enhances the catalytic activity of Pt/SnO_2 when incorporated into its formulation. Methods for depositing both promoted and unpromoted Pt/SnO_2 and Au/MnO_2 onto Cordierite monoliths have been developed. At the present time both the promoted catalyst composition and the monolith coating technology are proprietary. As shown in Figure 3, the activity of the Pt/SnO_2 is significantly enhanced by the addition of the promoter compound to the extent that catalytic activities at 35 °C are greater than the activities observed for the unpromoted cases at 55 °C. Additionally, the presence of high concentrations of CO_2 in the test gas has no deleterious effect on the activity, thus demonstrating that the promoted catalyst material should function in a laser gas environment. The promoted monolithic catalyst has already been successfully tested for one million pulses in a closed-cycle CO_2 laser. Initial tests on the Cordierite supported Au/MnO_2 showed a reduced activity when compared with its unsupported counterpart. It is believed that the activity was reduced by chloride contamination from $HAuCl_4$, the starting material used for gold impregnation. Alternatives to this contaminating material are under investigation.

Isotopic integrity studies were carried out on both Au/MnO_2 and promoted and unpromoted Pt/SnO_2. The catalysts were prereduced with H_2 and then reoxidized with $^{18}O_2$. Upon subsequent exposure to a stoichiometric test gas containing $C^{18}O$ and $^{18}O_2$ (Au/MnO_2 for 3 days and Pt/SnO_2 for 5 days), the isotopic composition of the CO_2 produced indicated no isotopic scrambling. The rare-isotope test gas mixture was also tested on the unlabeled Au/MnO_2 for 10 days and little isotopic scrambling was observed toward the end of the test indicating that surface labeling had occurred during the 10 day test. As shown in Figure 4, the isotope labeling process has little affect on the activity of the catalyst.

Figure 5 shows the evolution of the laser catalyst development effort at LaRC. As can be seen, the CO conversion rate (or pumping speed efficiency) increased from near zero standard milliliters CO per gram (sccm CO/g) of catalyst with the commercial catalyst tested at 35 °C in 1985 to approximately 3.5 sccm CO/g for the promoted monolith catalyst developed at LaRC in 1991.

Air purification studies have been initiated at LaRC for application of the most promising catalyst materials for removal of CO in purification of air. Figure 6 shows the results of comparative testing of 8-mesh Hopcalite granules, Au/MnO_2 powder, and promoted and unpromoted Pt/SnO_2 powders with a test gas of 500 parts-per-million by volume (ppm) CO in air. The benchmark material for CO removal has been Hopcalite for several decades. At both 35 °C and 24 °C the LaRC catalyst materials clearly were more efficient than Hopcalite for CO removal. At both temperatures the promoted and unpromoted Pt/SnO_2 powders completely removed CO from the gas stream. While the LaRC Cordierite monolith supported catalysts are yet to be tested for CO removal in an air medium, it is predicted that their performance will be considerably more efficient than their corresponding powder forms.

CONCLUSIONS

It is concluded from this effort that the promoted Pt/SnO_2 on a monolithic support functions well as both a CO-oxidation catalyst for use in closed-cycle CO_2 lasers at ambient laser gas temperatures and as a very effective means for quantitative removal of CO for application in the purification of air under ambient temperature conditions. Furthermore, the Au/MnO_2 catalyst, while limited in applications in which high concentrations of CO_2 are present, also functions well as a catalyst for the ambient temperature removal of CO in air. The Pt/SnO_2 catalysts could also be applicable in removal of CO and other incompletely combusted exhaust gas constituents in internal combustion engine exhausts during the cooler temperature exhaust gas condition which occurs during and shortly after cranking of a cold engine before exhaust gases heat up the catalytic converter to make it functional.

REFERENCES

1. "Closed-Cycle, Frequency Stable CO_2 Laser Technology," C. M. Batten, I. M. Miller, and G. M. Wood, Jr., eds., NASA Conference Publication, CP-2456, 1987.
2. Upchurch, B. T., et al., SPIE Proceedings, Vol. 1062 (1989).
3. Upchurch, B. T., et al., Proceedings of the International Conference on Lasers '89, 347-353 (1989).
4. Upchurch, B. T., et al., U.S. Patent Nos.: 4,829,035; 4,855,274, 4,839,330; 4,855,274; and 4,912,082.
5. Rogowski, R. S., et al., SPIE Proceedings, Vol. 415, 112-117 (1983).
6. Hess, R. V., et al., NASA TM 86415, April 1985.
7. Miller, I. M., et al., NASA TM 86421, April 1985.
8. Brown, K. G., et al., SPIE Proceedings, Vol. 663, 136-144 (1986).
9. Sidney, B. D., et al., SPIE Proceedings, Vol. 783, 162-167 (1987).
10. Hess, R. V., et al., SPIE Proceedings, Vol. 999 (1988).
11. "Low-Temperature CO-Oxidation Catalysts for Long-Life CO_2 Lasers," D. R. Schryer and G. B. Hoflund, eds., NASA Conference Publication, CP-3076, 1990.
12. Van Norman, John D., et al., "Low-Temperature CO-Oxidation Catalysts for Long-Life CO_2 Lasers," D. R. Schryer and G. B. Hoflund, eds., NASA Conference Publication, CP-3076, 181-191 (1990).
13. Upchurch, B. T., et al., SPIE Proceedings, Vol. 1416, 21-29 (1991).
14. Hoflund, G. B., et al., SPIE Proceedings, Vol. 1062 (1989).

15. Hoflund, G. B., et al., Thin Solid Films, Vol. 169, 69-77 (1989).
16. Gardner, S. D., et al., J. Catalysis, Vol. 115, 132-137 (1989).
17. Gardner, S. D., et al., J. Catalysis (in press).
18. Gardner, S. D., et al., SPIE Proceedings, Vol. 1062, 21-28 (1989).
19. Schryer, D. R., et al., J. Catalysis, Vol. 122, 193-197 (1990)
20. Drawdy, J. E., et al., Surface and Interface Analysis, Vol. 16, 369-374 (1990).
21. Gardner, S. D., et al., J. Phys. Chem., Vol. 95, 835-838 (1991).
22. Gardner, S. D., et al., J. Catalysis, Vol. 129, 114-120 (1991).
23. Schryer, D. R., et al., J. Catalysis, Vol. 130, 314-317 (1991).

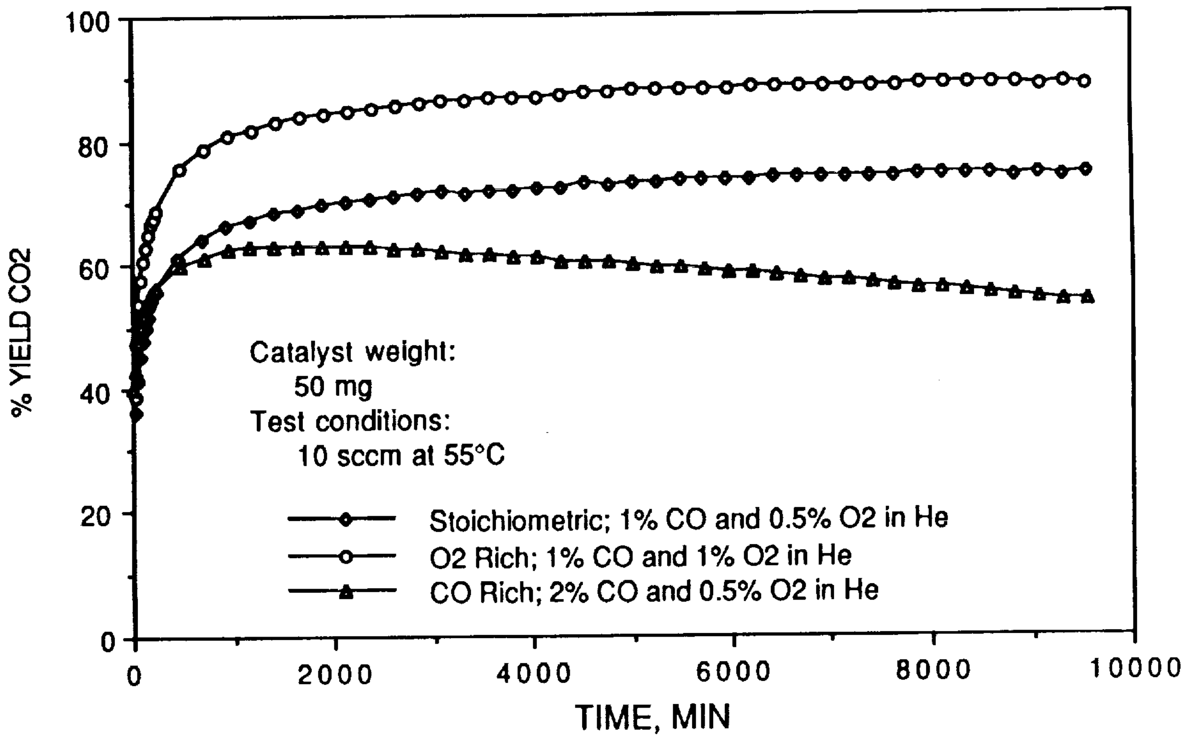


Figure 1. Effect of test gas on the activity of Au/MnOx.

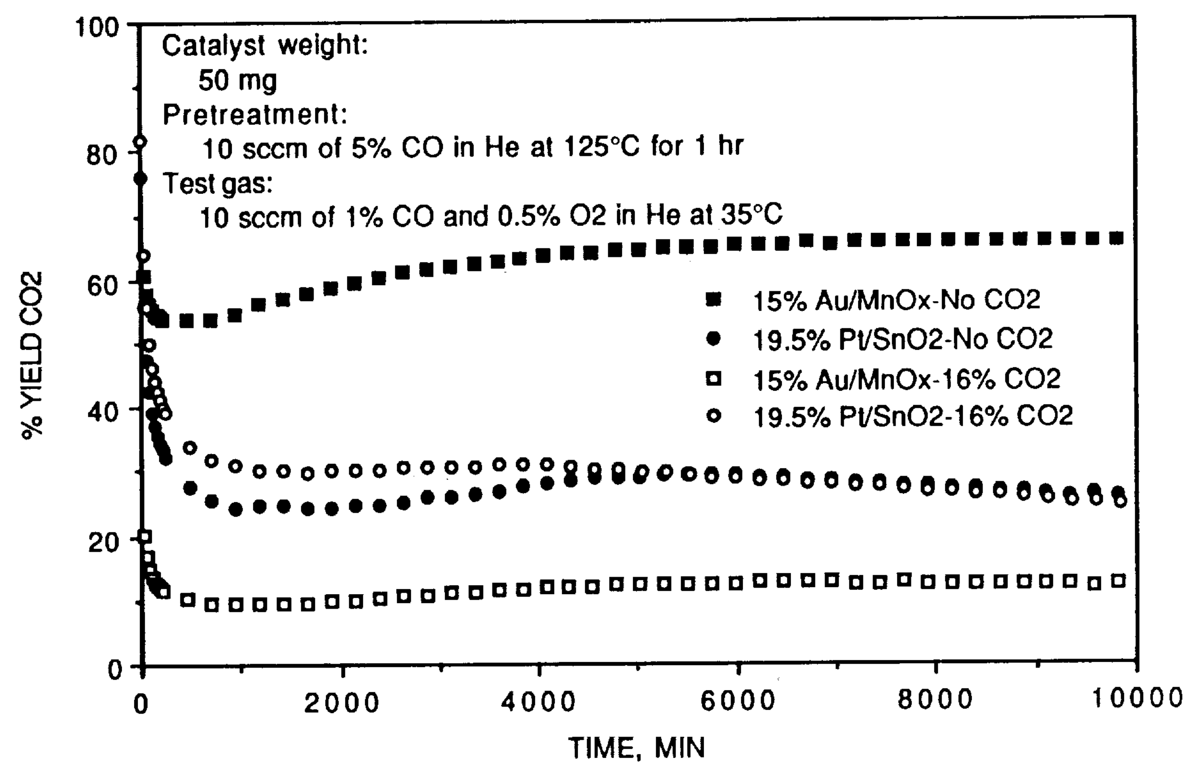


Figure 2. Effect of CO2 rich test gas on the activity of Au/MnOx and Pt/SnO2 catalysts.

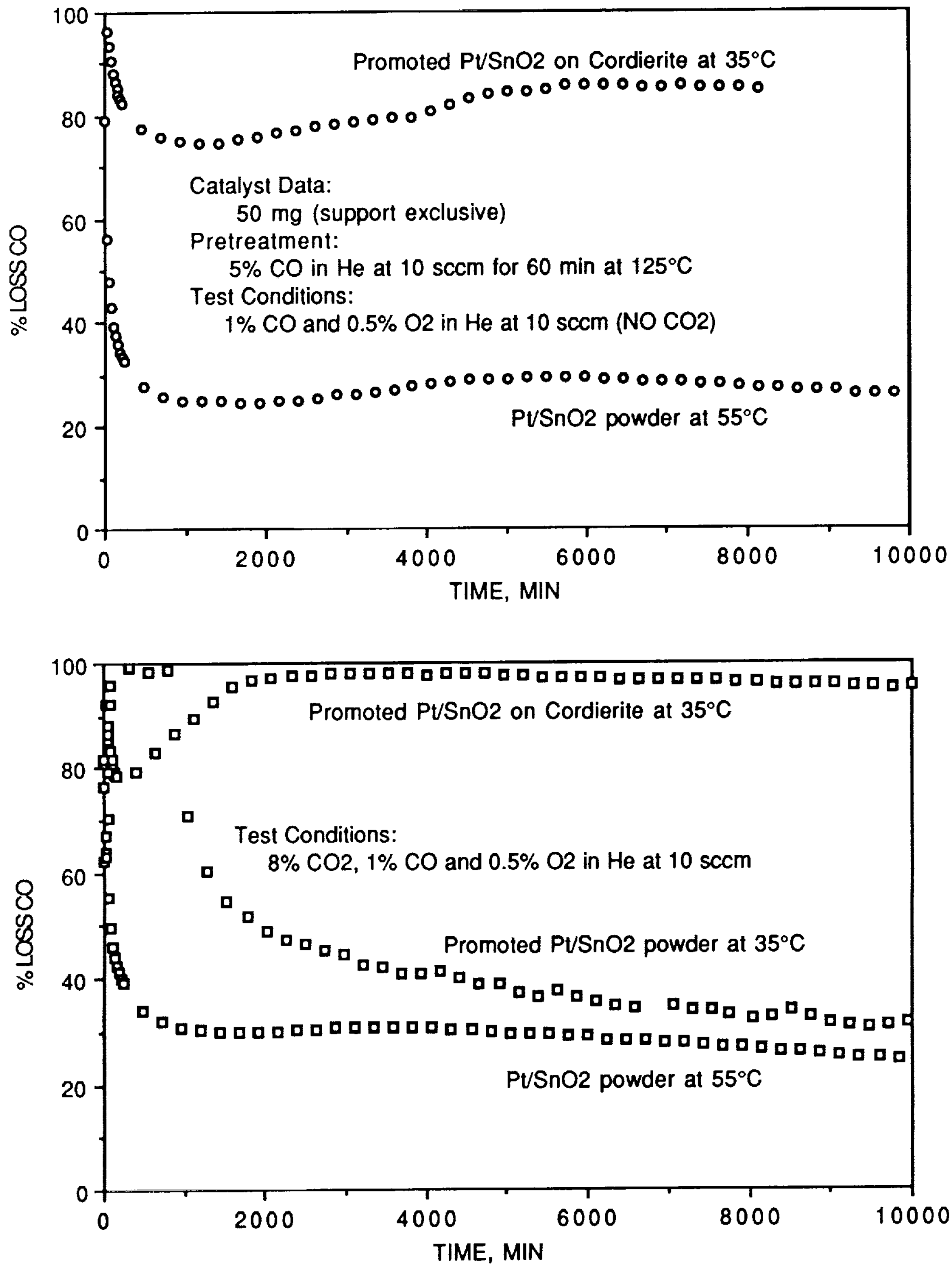


Figure 3. Effect of CO_2 in test on the activity of promoted and unpromoted Pt/SnO_2.

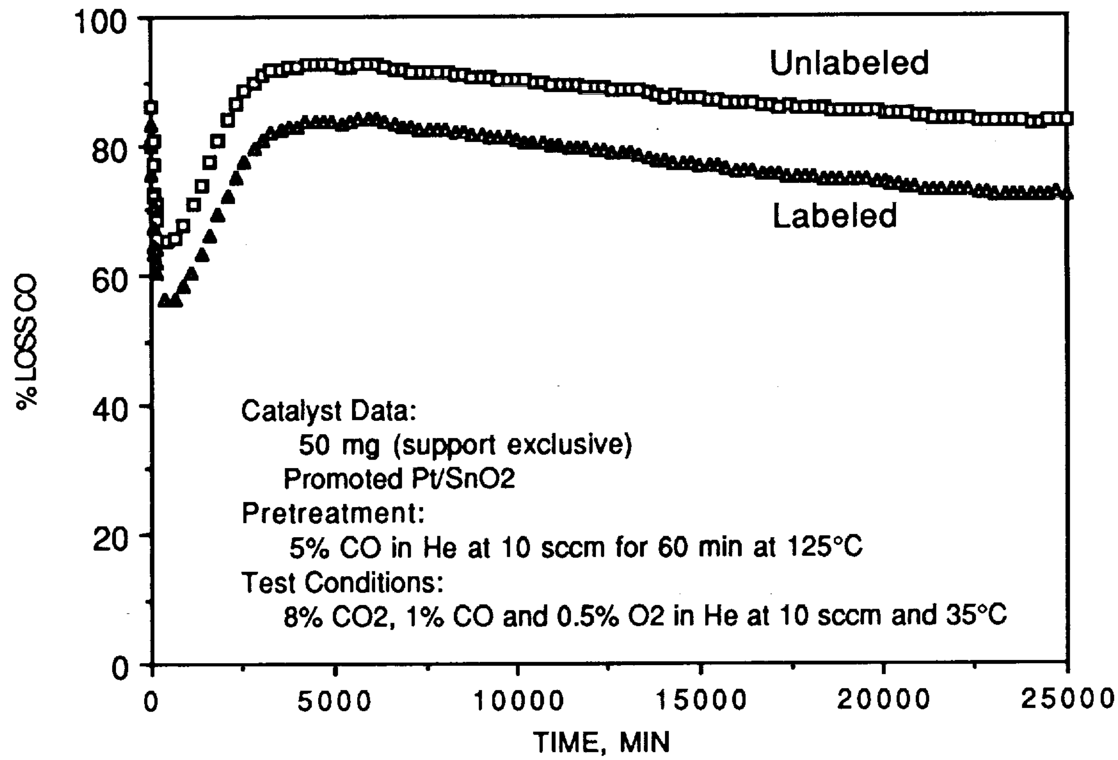


Figure 4. Effect of O-18 labeling on the activity of promoted Pt/SnO2 on Cordierite.

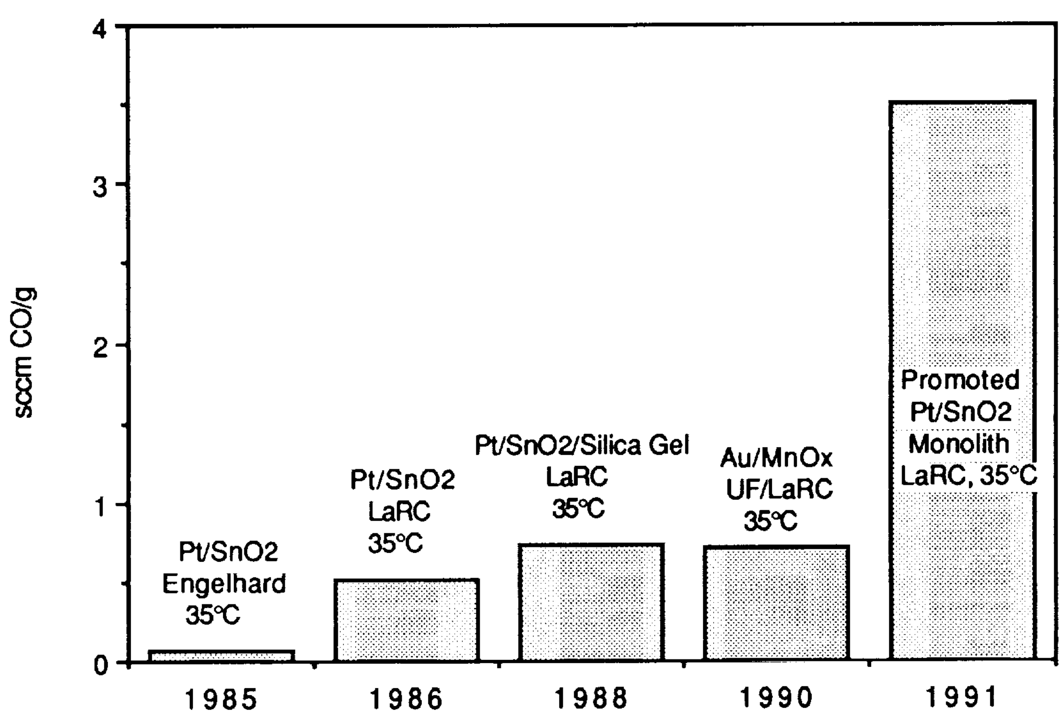


Figure 5. Catalyst efficiencies in laser test gas.

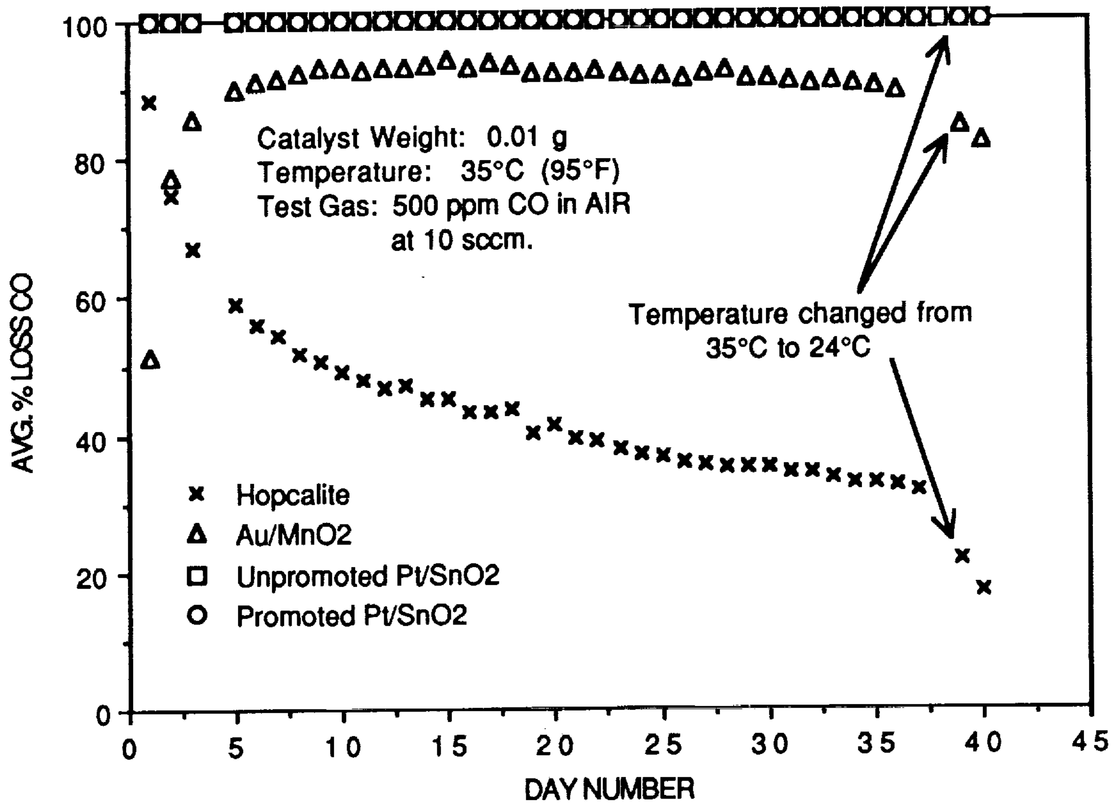


Figure 6. Daily average comparison of CO-oxidation catalysts in air.

MATERIALS SCIENCE

(Session B4/Room A2)

Wednesday December 4, 1991

- High-Temperature Adhesives
- Fluorinated Epoxy Resins with High Glass Transition Temperatures
- Polyimides Containing Pendent Siloxane Groups
- Corrosion-Protective Coatings From Electrically-Conducting Polymers

N92-22460

HIGH TEMPERATURE ADHESIVES

Terry L. St. Clair, Head
Polymeric Materials Branch
NASA Langley Research Center
Hampton, VA 23665-5225

ABSTRACT

The aerospace and electronics industries have an ever increasing need for higher performance materials. In recent years linear aromatic polyimides have been proven to be a superior class of materials for various applications in these industries. The use of this class of polymers as adhesives is continuing to increase. Several NASA Langley-developed polyimides show considerable promise as adhesives because of their high glass transition temperatures, thermal stability, resistance to solvents/water, and their potential for cost-effective manufacture.

INTRODUCTION

Over the past 2 decades several commercially attractive polyimide adhesives have been developed at the NASA Langley Research Center [1-4]. These materials were developed as structural adhesives for use in the 200-300°C range, however they appear to have utility for other end-use applications. One particular adhesive, LARC-TPI (Langley Research Center Thermoplastic Polyimide), has become commercially available. Three other linear thermoplastic polyimides have been developed in more recent years that exhibit characteristics that for one reason or another have made them candidates for scaleup by NASA. They are LARC-CPI (Crystalline Polyimide), LARC-ITPI (Isomeric TPI) and LARC-IA (Improved Adhesive). Their properties will be discussed.

LARC-CPI

This semicrystalline poly(keto-ether-imide) can be readily processed above its crystalline melt temperature (343°C) to form adhesive bonds of very high strength [2]. The structure of this polyimide as well as the other three mentioned in the introduction are shown in Figure 1. After the bonding operation, the adhesive crystallizes upon cool down and can continue to crystallize during thermal exposures near the glass transition temperature of the polymer (223°C). Figure 2 illustrates this behavior. The adhesive performance of LARC-CPI is shown in Figure 3. The tendency of this adhesive to gain strength when aged at 232°C and tested at this same temperature is attributed to the development of crystallinity.

The resistance to organic solvents and to base hydrolysis that is exhibited by LARC-CPI is evidently due to the crystallinity. In 20 percent sodium hydroxide solution, LARC-CPI shows no tendency to hydrolyze even after a week period. This is a phenomenal property that allows this material to be used in some very hostile environments that would totally hydrolyze just about all other polyimides. LARC-CPI, again because of its crystallinity, has an extremely low level of water pickup (less than 0.5 percent).

LARC-ITPI

LARC-ITPI was developed as a cost effective alternative to LARC-TPI. A major drawback to the large scale commercialization of LARC-TPI has been associated with its cost which is primarily caused by the diamine component, 3,3'-diaminobenzophenone. The diamine that is used in LARC-ITPI is the very reasonably priced meta-phenylenediamine (MPD). The adhesive performance of LARC-ITPI is shown in Figure 4. The IDPA-MPD is the LARC-ITPI with two versions being represented. The center set of bar graphs represents the controlled molecular weight form of the polymer which has four percent endcap (EC 4.00). Its adhesive performance is quite comparable to LARC-TPI. Their glass transition temperatures are similar (258°C and 260°C). These two adhesives exhibit very similar properties after 1000 hours of exposure at 232°C.

LARC-IA

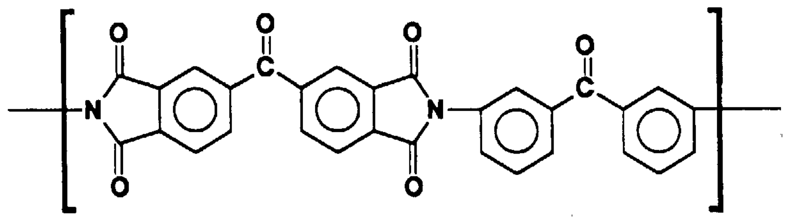
LARC-IA is the designation for the polyimide based on the novel diamine 3,4'-oxidianiline (3,4'-ODA). This diamine became available due to the development of a novel high tensile-strength, high-modulus polyamide in Japan. The use of this diamine in a BTDA-based polyimide resulted in an adhesive with a glass transition temperature of 243°C. Just as with the LARC-ITPI it was necessary to control the molecular weight this time with five percent endcapper (phthalic anhydride). Some selected adhesive properties of LARC-IA are shown in Figures 5 and 6. In these Figures LARC-IA is compared with and without aluminum powder which improves its high temperature performance.

SUMMARY

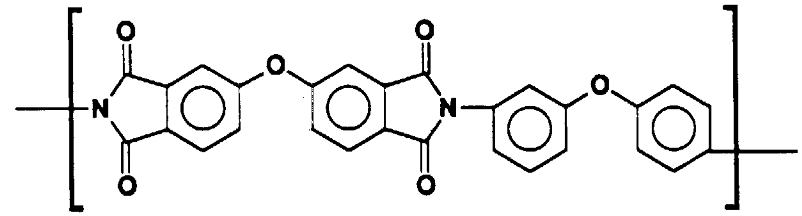
Several high temperature polymers have been developed at NASA Langley Research Center that have been shown to exhibit exceptional adhesive properties. The four systems that have been investigated the most are the commercially available LARC-TPI and the three experimental systems designated LARC-CPI, LARC-ITPI and LARC-IA. Each of these materials has special attractive properties which make them commercially attractive.

REFERENCES

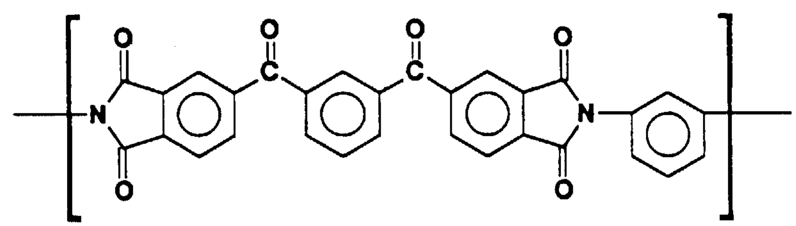
1. St. Clair, A. K. and St. Clair, T. L., SAMPE Quarterly, 13(1), pp. 20-25 (1981).
2. Hergenrother, P. M. and Havens, S. J., Inter. SAMPE Symp., 36, pp. 56-63 (1991).
3. Pratt, J. R.; St. Clair, T. L. and Progar, D. J., U.S. Patent 4,937,317, June 1990.
4. Progar, D. J. and St. Clair, T. L., J. Adhesion Sci. Technology, 4(7), pp. 527-549 (1990).



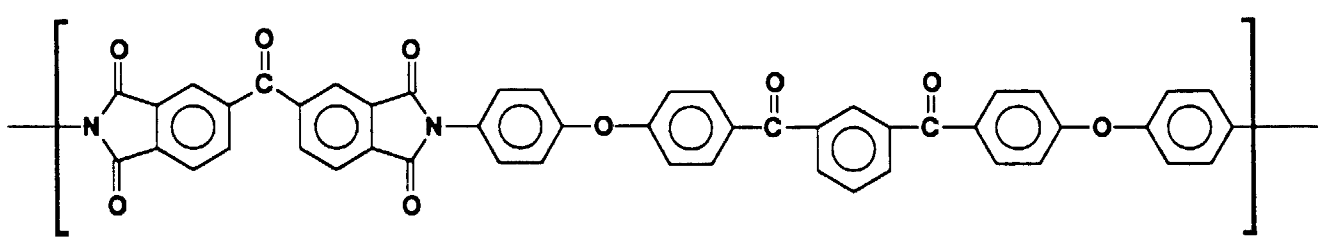
LARC-TPI



LARC-IA



LARC-ITPI



LARC-CPI

FIGURE 1 POLYIMIDE ADHESIVES

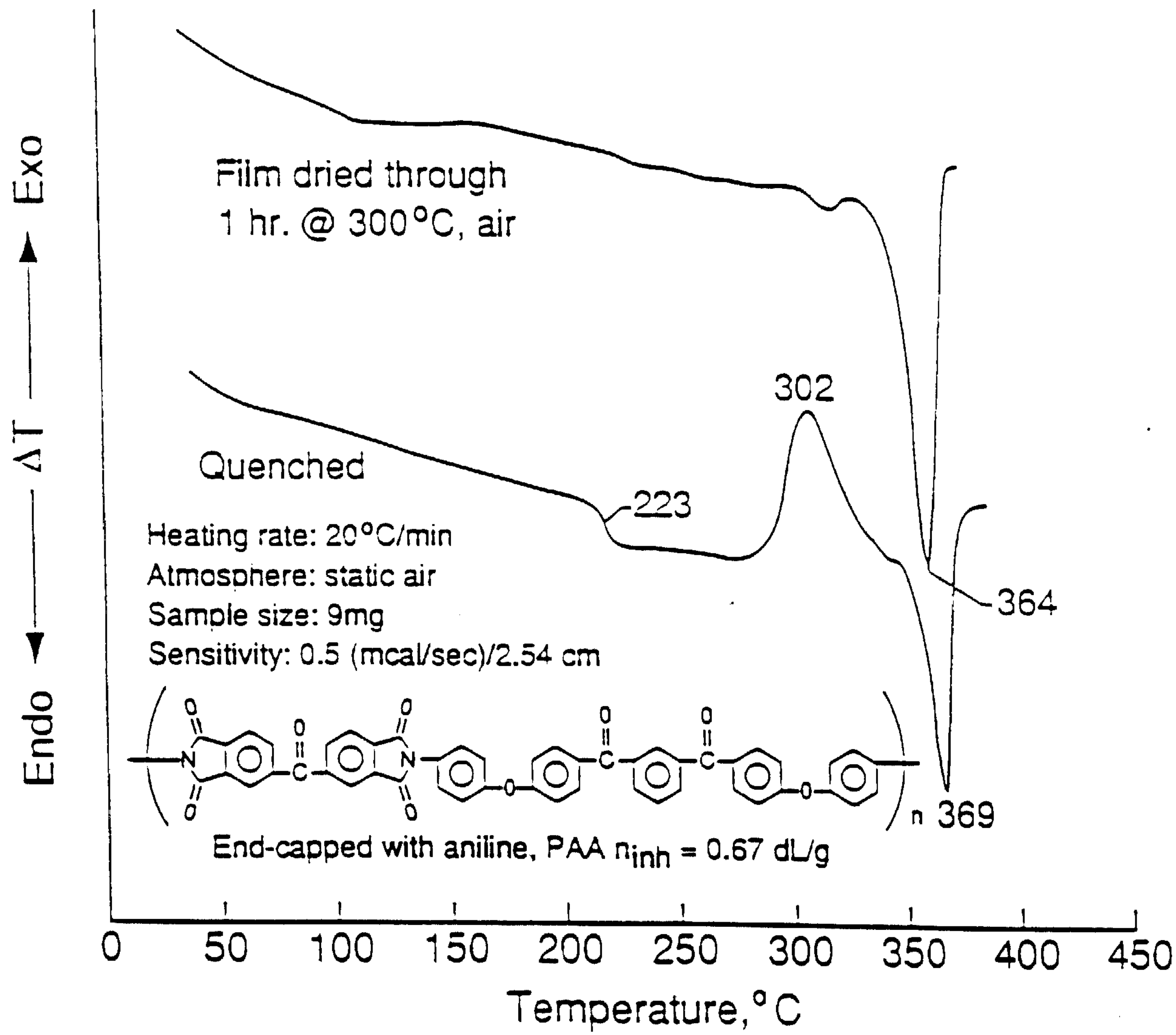


FIGURE 2 DIFFERENTIAL SCANNING CALORIMETRIC CURVES - LARC-CPI

TEST CONDITION	STRENGTH, PSI	FAILURE
25°C	6250	> 95% COHESIVE
25°C AFTER 3-DAY WATER BOIL	5140	~ 90% COHESIVE
25°C AFTER 72 HR HYDRAULIC FLUID SOAK	5590	~ 70% COHESIVE
25°C AFTER 1000 HR @ 232°C	7120	~ 100% COHESIVE
25°C AFTER 5 HR @ 300°C, 100 PSI	6130	> 95% COHESIVE
25°C AFTER 100 HR @ 316°C	4590	~ 70% COHESIVE
177°C	4510	> 95% COHESIVE
177°C AFTER 4 HR @ 300°C, 100 PSI	4690	~ 100% COHESIVE
232°C	590	~ 95% ADHESIVE
232°C AFTER 100 HR @ 232°C	1840	~ 50% COHESIVE
232°C AFTER 1000 HR @ 232°C	2740	~ 50% COHESIVE
232°C AFTER 5 HR @ 300°C, 100 PSI	2800	~ 80% COHESIVE
232°C AFTER 100 HR @ 316°C	3670	> 95% COHESIVE

*PASA-JELL 107 SURFACE TREATMENT; INHERENT VISCOSITY OF POLY(AMIC ACID) = 0.50 dL/g; BONDING CONDITIONS, 400°C, 1000 PSI, 15 MIN; 112 E-GLASS TAPE CONTAINED 0.1% VOLATILES, BONDLINE THICKNESS 5-6 MILS

FIGURE 3 LARC-CPI ADHESIVE DATA

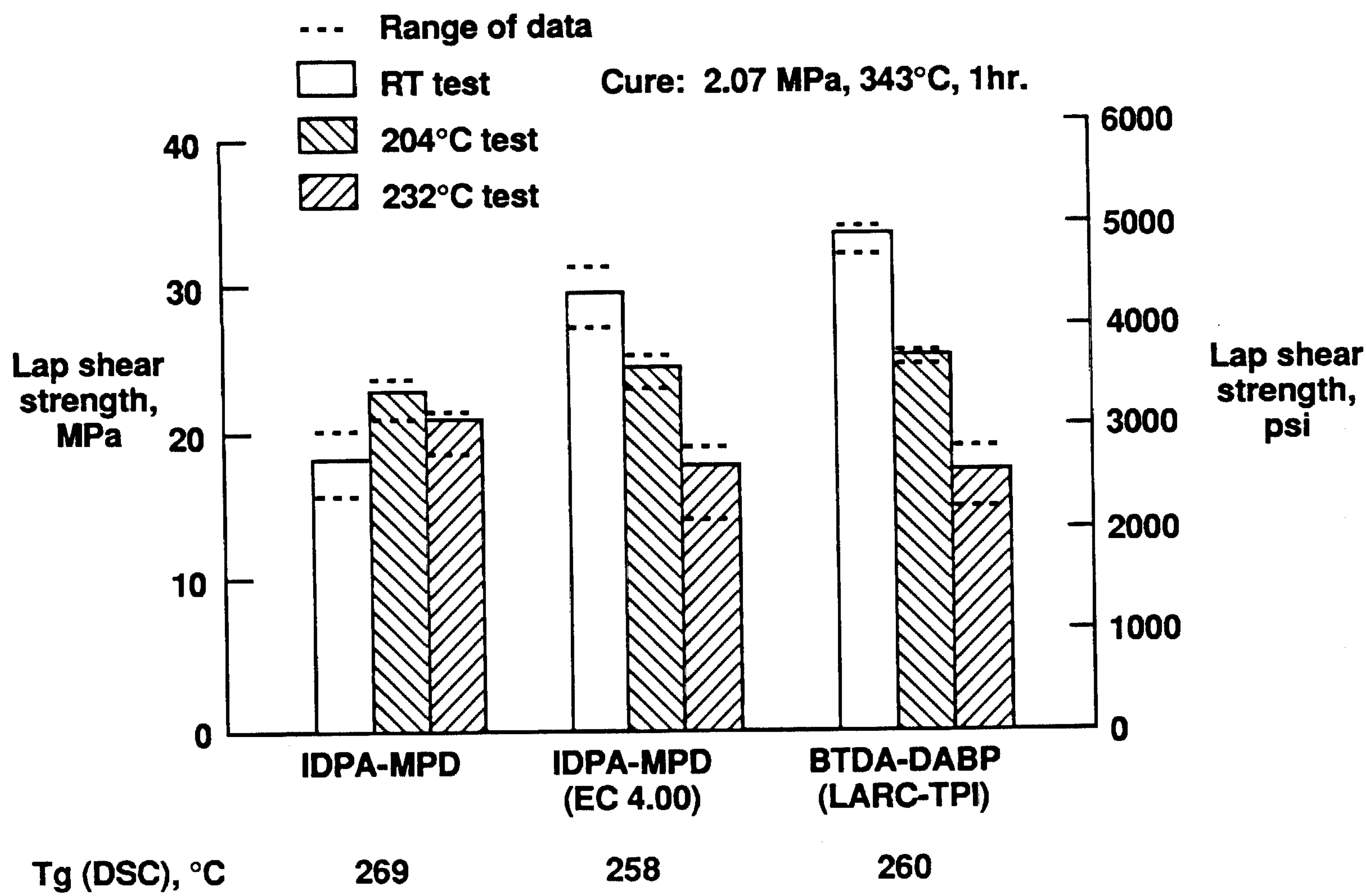


FIGURE 4 LARC–ITPI ADHESIVE DATA

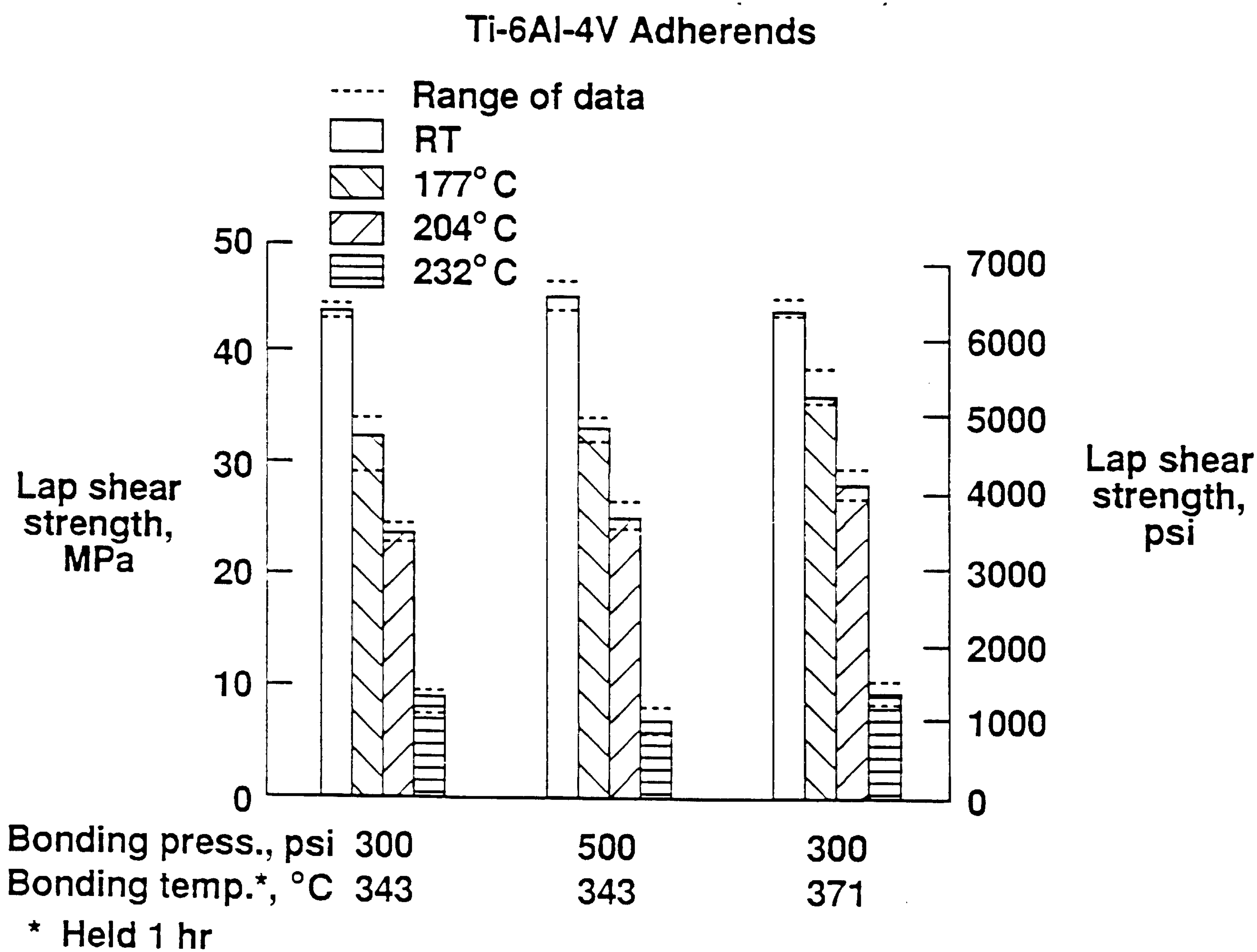


FIGURE 5 ADHESIVE DATA FOR LARC–IA

Adhesive	FWT[a] [MPa (psi)]	G_{1c}[b] $\left[J/m^2 \left(\frac{in. lb}{in.^2} \right) \right]$
LARC-IA (5% PA)	1.7 (247)	560 (3.2)
	2.6 (378)	438 (2.5)
LARC-IA (5% PA) 50% Al powder	2.3 (340)	298 (1.7)
	2.4 (355)	210 (1.2)

[a] Specimens were fabricated from Ti-6Al-4V facesheets [0.05 cm (0.020 in.)] and Ti core [0.95 cm (0.37 in.) cell size] given a Pasa Jell 107 surface treatment before priming and bonding

[b] Specimens were fabricated from Ti-6Al-4V 0.13 cm (0.050 in.) thick and given a Pasa Jell 107 surface treatment before priming and bonding

FIGURE 6 FLATWISE TENSILE STRENGTH AND FRACTURE ENERGY FOR LARC–IA

N92-22461

FLUORINATED EPOXY RESINS WITH HIGH GLASS TRANSITION TEMPERATURES

James R. Griffith
Naval Research Laboratory
Washington, DC 20375-5000

ABSTRACT

Easily-processed liquid resins of low dielectric constants and high glass transition temperatures are useful for the manufacture of certain composite electronic boards. That combination of properties is difficult to acquire when dielectric constants are below 2.5, glass transition temperature are above 200°C and processability is of conventional practicality. A recently issued patent (U.S. 4,981,941 of Jan. 1, 1991) teaches practical materials and are the culmination of 23 years of research effort and 15 patents owned by the Navy in the field of fluorinated resins of several classes. In addition to high fluorine content, practical utility has been emphasized.

INTRODUCTION

The lowest dielectric constants obtainable with solid polymeric materials are just below 2.0, and these are fluoropolymers which have inconvenient processing characteristics for the production of composite structures. A combination of properties which include dielectric constants around 2.5 $\pm$ 0.2, convenient processability due to neat liquid solidification, and glass transition temperatures above 200°C can now be obtained with fluorinated epoxy resins and mixed fluoroanhydride curing agents. Commercial facilities for the production of these resins in quantity exist and await sufficient demand to provide the potential markets. The most obvious and immediate market is the electronic printed circuit or composite component board manufacture in which the low dielectric constants would enhance performance and the high glass transition temperatures would aid construction.

MATERIALS

Fluoroepoxy Resins

The synthesis of heavily fluorinated epoxy resins with convenient processing characteristics was undertaken at the Naval Research Laboratory in 1968 and effective materials were in hand in the early 1970's (1,2). Several basic materials patents were generated regarding the fluoroepoxies (3,4) as well as fluoropolyurethanes (5) and fluoroacrylics (6). In this presentation the discussion will be confined to the fluoroepoxies and means of producing high glass transition temperature versions which is accomplished by the proper selection of curing agents and processes.

Curing Agents

Since the dielectric constant of a polymer is roughly inversely proportional to the fluorocarbon content, it is desireable to have fluorine in the curing agent as well as the resin. However, the glass transition temperature also has an inverse relationship to the fluorine content and since the dielectric constant and glass transition temperature both fall with increasing fluorine, it is necessary to offset the declining glass transition temperature with an additional structural factor. An effective factor for this is the crosslink density which is controlled largely by the functionality of the curing agent. The glass transition temperature increases with increasing crosslink density while the dielectric constant is not affected.

These considerations suggest anhydride curing agents as the materials of choice and we have patented several fluorinated versions (7). There is at least one fluorinated dianhydride that is commercially available, and the dianhydrides are particularly effective for raising glass transition temperatures. However, they always have very high melting points, are usually of limited solubility in the resins, and cause premature gelation of the resin system if forced into compatibility by heating. On the other hand, the monoanhydrides, including fluoro varieties, are often low melting, convenient materials. The dianhydrides will often dissolve to a limited extent in the molten monoanhydrides, and this fact offers a practical means by which a relatively high glass transition resin can be obtained. Thus, the dianhydride is dissolved into the monoanhydride and upon sudden cooling, a glass is obtained which may be dissolved at will in the resin of choice. When the lowest possible dielectric constants are required, all of the components may have fluorocarbon in the molecular structures.

Catalysts

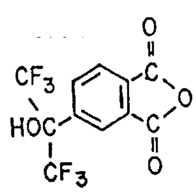
Anhydride-cured epoxy resins, whether fluorinated or not, require minimum cure temperatures of about 80°C and it is common that final temperatures of about 150°C are employed. The gel times of such systems are strongly influenced by the type and quantity of catalyst used and the quarternary ammonium salts are often the catalysts of choice. Very small quantities (about 0.1% of resin mass) are commonly used and small changes in quantity can have large reaction rate effects. For a practical manufacturing technique the exact compositions of the resin must be derived empirically although this is not a difficult, or impossible, determination when the mixed monoanhydride-dianhydride system is employed.

Solvents

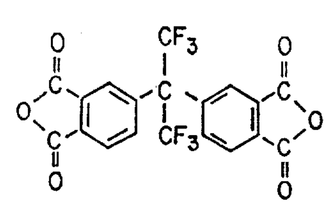
It is not necessary to employ a solvent in the production of a fiber filled composite or circuit board since the neat liquids are often of sufficiently low viscosity to infiltrate effectively. Because of the fluorocarbon content they also have low surface tensions and are thus excellent wetting fluids. However, if the gelation times are too short for a given manufacturing technique, or the viscosities are too high, it is possible to employ a solvent. Ketones such as methyl ethyl ketone or acetone are often the solvents of choice. It is also possible to employ the "prepreg" technique since the impregnated systems are relatively stable at room temperature or below and this technique makes the elimination of the volatile solvent relatively easy.

CHEMICAL FORMULAS

The following structural formulas identify the more important materials available, or potentially available:

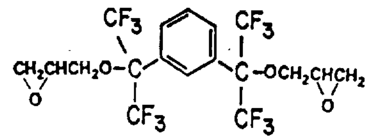


"MONO" ANHYDRIDE

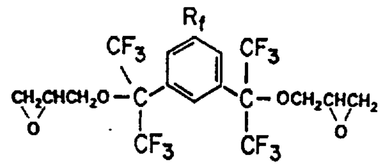


"DI" ANHYDRIDE

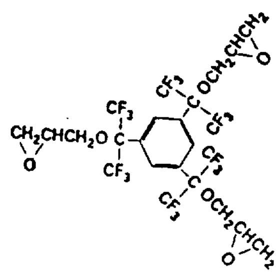
The highest glass transition temperatures are obtained with the more highly functional components. For example, the highest value obtained to date is 249°C which resulted from a composition of the trifunctional epoxy rich in the dianhydride. It is probable that even higher glass transition temperatures can be obtained, approaching 300°C, but the processing difficulties involved in reaching such levels may be formidable.



"C-O" FLUOROEPOXY



"C-6" IF Rf IS PERFLUOROHEXYL, FOR EXAMPLE



"TRI"FUNCTIONAL FLUOROEPOXY

REFERENCES

1. Reines, S. A., Griffith, J. R. and O'Rear, J. G., J. Org. Chem. 36, 1209 (1971).
2. Griffith, J. R., O'Rear, J. G. and Reines, S. A., CHEMTECH, 311-316, May 1972.
3. U. S. Patent 3,549,591, Dec. 22, 1970.
4. U. S. Patent 3,879,430, Apr. 22, 1975.
5. U. S. Patent 4,132,681, Jan. 2, 1979.
6. U. S. Patent 4,284,747, Aug. 18, 1981.
7. U. S. Patent 4,045,408, Aug. 30, 1977.

N92-22462

POLYIMIDES CONTAINING PENDENT SILOXANE GROUPS

John W. Connell
NASA Langley Research Center
Hampton, VA 23665-5225

ABSTRACT

The incorporation of siloxane units into the backbone of aromatic polyimides has been shown to impart certain advantages over the unmodified polyimides. These include enhanced solubility, lower moisture adsorption, lower dielectric constant, improved toughness and surface modification. In addition, when exposed to an atomic oxygen environment these materials form an in-situ silicate (SiO_2) surface coating which protects the underlying material from further erosion. These unique advantages make polyimide-siloxanes useful in a variety of electronic and aerospace applications. For example, these materials find use in the microelectronic industry as interlayer dielectrics where thermal stability, low dielectric constant, low water adsorption and good adhesion to various substrates is required. Materials of this type may find use as films and coatings in a space environment where resistance to atomic oxygen erosion is needed.

As part of an effort on high performance polymeric materials for potential aerospace applications, polyimides containing pendent siloxane groups are under investigation. These materials were prepared by reacting a functionalized siloxane compound with polyimides containing benzhydrol groups. Thin films of the polymers exhibited glass transition temperatures ranging from 167 to 235°C. Tensile strengths and moduli measured at 23°C ranged from 11-14 ksi and 250-450 ksi, respectively. The dielectric constant was lowered substantially from that of the unmodified polyimide. Preliminary data after exposure to a simulated atomic oxygen environment (Asher) indicates that these materials form a silicate surface coating. The chemistry, physical and mechanical properties of these materials as well as potential applications will be discussed.

INTRODUCTION

Organic polymeric materials are currently being considered for long term use (> 10 years) in structural (adhesives and composites) and functional (films and coatings) applications on spacecraft. Polymeric materials offer attractive features such as low density, low thermal expansion over a relatively large temperature range, high strength and stiffness. In addition, they can provide unique performance in specialized applications where optical transparency, surface smoothness or adhesion is of critical importance. Although organic polymeric materials have been utilized successfully on spacecraft for short term missions, the long term durability of these materials in space is uncertain. Of particular concern is the durability of polymeric materials in low earth orbit (LEO) where atomic oxygen (AO) is prevalent.. Due to the high erosion rates of polymeric materials by atomic oxygen, they must be protected by a surface coating. Coatings that appear to eliminate or substantially reduce erosion of polymeric materials by AO include aluminum oxide[1], silicon dioxide[1], polytetrafluoroethylene[2], chromium oxide[2], copper-sapphire[2] and indium-tin oxide[3]. To be effective the coatings must be uniform, ~500-2000Å thick and pinhole and defect free.

A number of literature reports concerning polyimides containing siloxane groups in the mainchain (i. e., polymer backbone) are available. A U. S. patent describing polyimides of this type dates back to 1961[4]. The first report in the open literature appeared in 1966[5]. Since then a number of literature reports[6-11] and U. S. patents[12-17] concerning polyimides containing siloxane groups in the backbone have been disclosed. Recently, a report on polyimides containing pendent siloxane groups was published[18].

As part of a NASA effort to develop materials technology for potential space applications, a series of polyimides containing pendent siloxane groups were synthesized, characterized and evaluated under a simulated atomic oxygen environment using a radio frequency generated oxygen plasma asher. The oxygen plasma asher contains some species that are not present in LEO and therefore does not accurately or adequately represent that orbital environment. Also, other types of radiation (i. e., ultraviolet, electron and proton) and temperature cycling are present in LEO and may cause synergistic effects resulting in accelerated degradation. However, the device is useful for testing the hypothesis that the polyimides containing pendent siloxane groups will form an in-situ SiO_2 surfacing coating during exposure. Pendent siloxane groups should be advantageous as compared to backbone siloxane groups since cleavage of the organo-silicon bonds would not necessarily result in molecular weight degradation (i. e., backbone cleavage). The results of this study are presented herein.

EXPERIMENTAL

The polyimides containing pendent siloxane groups were prepared from polyimides containing benzhydrol groups. One polyimide containing benzhydrol groups used in this study is commercially available and the other was an experimental material that was synthesized in-house. A detailed representative experimental procedure is given below for the preparation of the polyimides containing pendent siloxane groups from commercial and experimental polyimides containing benzhydrol groups.

Experimental Polyimide Containing Benzhydrol Groups

Into a 100 ml three neck round bottom flask equipped with a mechanical stirrer, nitrogen gas inlet and drying tube filled with calcium carbonate was placed 3,3'-diaminobenzhydrol (3.3210g, 15.5 mmol) and DMAc (15.0 ml). The mixture was stirred at 23°C until the diamine dissolved (~15 minutes). To this solution was added 3,3',4,4'-benzophenonetetracarboxylic dianhydride (4.9944g, 15.5 mmol) and DMAc (10.0 ml) to give a final concentration of 25.0% solids. The solution was stirred at 23°C for ~16 hr under nitrogen to give a viscous poly(amide-acid) solution (inherent viscosity of a 0.5% solution in DMAc at 25°C was 0.81 dL/g). The poly(amide-acid) solution was diluted to 15% solids by the addition of DMAc (8 ml) and transferred to a pressure equalizing addition funnel which had previously been flushed with nitrogen.

Into a 250 ml three neck round bottom flask equipped with a mechanical stirrer, thermometer, nitrogen gas inlet, Dean Stark trap and reflux condenser was placed DMAc (20 ml) and xylenes (30 ml). The liquids are heated to reflux (~150°C) and maintained for ~5 hr. The poly(amide-acid) solution was subsequently added dropwise to the refluxing DMAc/xylenes mixture over ~1 hr period. Refluxing was continued for 1 hr after all of the poly(amide-acid) solution had been added. The xylenes was subsequently removed via the Dean Stark trap and the polyimide was precipitated into water in a high speed blender. The polymer was washed repeatedly in water and dried at 150°C for ~4 hr under vacuum. The polyimide had a glass transition temperature (T_g) of 267°C and an inherent viscosity of 0.43 dL/g measured on a 0.5% solution in DMAc at 25°C.

Polyimide Containing Pendent Siloxane Groups

Into a 100 ml three neck round bottom flask equipped with a mechanical stirrer, nitrogen gas inlet and pressure equalizing addition funnel was placed the previously described polyimide containing benzhydrol groups (1.56g, 3.1 mmol based on hydroxy group content, assuming a molecular weight of 20,000 g/mole) and DMAc (9 ml, 15% solids). The mixture was stirred at 23°C until the polyimide dissolved (~1 hr) and platinic acid (55 mg) was subsequently added. Into the pressure equalizing addition funnel was placed 1,1,2,2,3,3,3-heptamethyl-trisiloxane (0.73g, 3.25 mmol) and toluene (5 ml). The siloxane dissolved rapidly in the toluene and the solution was subsequently added dropwise to the polymer solution over a 30 minute period. The solution was stirred at 23°C for 16 hr, filtered through 5.0 micron filter paper under ~20 psi and cast into a thin film. The film was stage-dried to 235°C and held for 1 hr at 235°C under vacuum. The translucent orange film exhibited a T_g of 235°C. Infrared spectroscopic analysis of the film indicated that the reaction had proceeded as anticipated. Tensile strength and modulus of thin film specimens at 23°C of 12.5 and 391 ksi, respectively were obtained.

Polyimide Containing Pendent Siloxane Groups From Commercial Polyimide

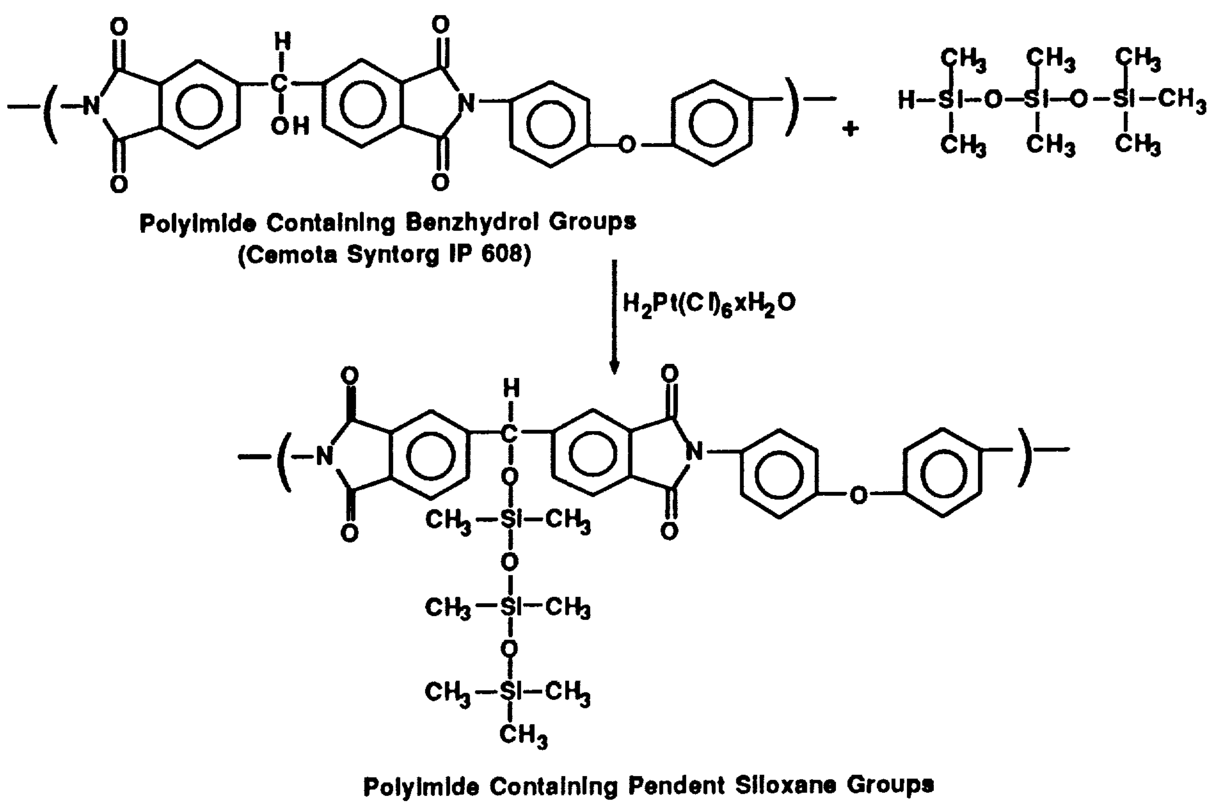
Into a 100 ml three neck round bottom flask equipped with a mechanical stirrer, nitrogen gas inlet, and pressure equalizing addition funnel was placed Cemota Syntorg IP 608 polyimide containing benzhydrol groups (10.12g, 40.2 mmol based on hydroxy group content, assuming a molecular weight of 36,000 g/mole) and N-methyl-2-pyrrolidinone (NMP) (58 ml, 15% solids). The mixture was stirred at 23°C under nitrogen until the polymer had dissolved (~1 hr) and platinic acid (105 mg) was subsequently added. Into the pressure equalizing addition funnel was placed 1,1,2,2,3,3,3-heptamethyltrisiloxane (9.86g, 44.3 mmol) and toluene (25 ml). The siloxane solution was added to the polyimide solution dropwise over a 1 hr period. The solution was stirred at 23°C under nitrogen for 16 hr and subsequently filtered through a 5.0 micron filter under pressure (~20 psi). A thin film was cast from the solution onto plate glass. The film was dried to a tack-free state in a dust-proof chamber and stage-dried to 225°C and held at 225°C for 1 hr under vacuum. The translucent yellow/green film exhibited a T_g of 167°C. Tensile strength and modulus of thin film specimens at 23°C of 11 and 250 ksi, respectively were obtained.

Asher Exposure

Simulated atomic oxygen exposures were performed on thin films (0.5 x 0.5 in., ~1-3 mils thick) of the polyimides containing pendent siloxane groups in a Tegal Plasmod Asher. The Asher was operated at 500 millitorr, 100 Watts of radio frequency, O_2 pressure of 3 psi and a flow rate of 50 standard cubic centimeters per minute. Since the Asher was not calibrated, simultaneous exposures of Kapton® and Ultem® were performed with each experimental polyimide containing pendent siloxane groups. The Kapton® and Ultem® films served as standards allowing for direct comparison with the films of the polyimides containing pendent siloxane groups. Exposures were performed for up to 8 hours and the weight loss of the films were monitored as a function of exposure time.

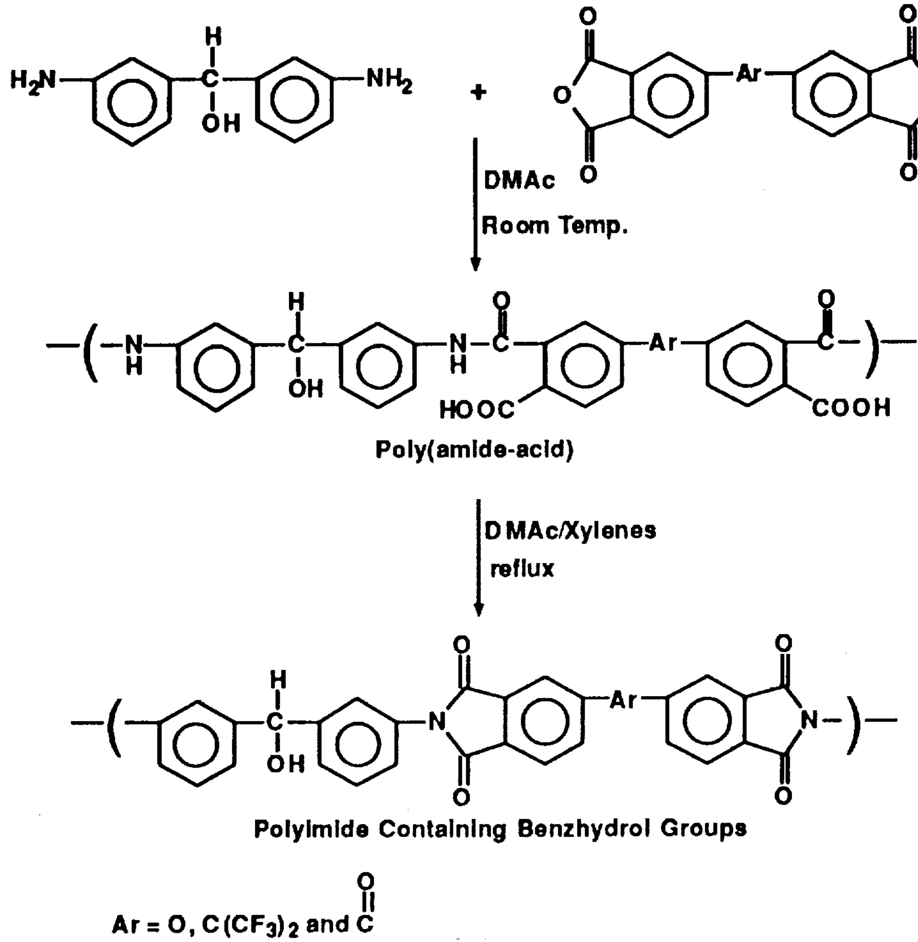
RESULTS AND DICUSSION

The polyimides containing pendent siloxane groups were prepared by reacting polyimides containing benzhydrol groups with heptamethyltrisiloxane as shown in equation 1. The reaction solution was subsequently used to cast thin films. After stage-drying the polymer films up to ~ 225-250°C to remove residual solvent the films were characterized by infrared spectroscopy, differential scanning calorimetry (DSC), thermogravimetric analysis (TGA) and measurement of thin film tensile properties.



Equation 1. Synthesis of polyimides containing siloxane groups.

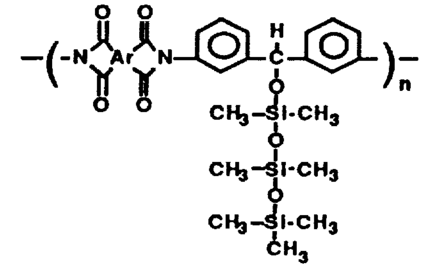
The synthesis of the experimental polyimides containing benzhydrol groups is depicted in equation 2. After isolation and characterization these polyimides were reacted with the heptamethyltrisiloxane as shown in equation 1.

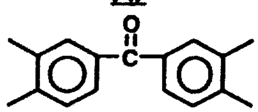
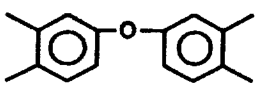
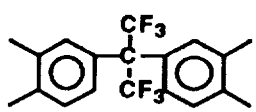


Equation 2. Synthesis of Polyimides Containing Benzhydrol Groups.

Polymer characterization is shown in Table 1 for the polyimides containing pendent siloxane groups prepared from the experimental benzhydrol containing polyimides. All of these films were translucent which is indicative of some degree of phase separation, although only one T_g was detected by DSC. As expected the T_gs of the polyimides containing siloxane groups are significantly lower than the corresponding polyimides containing benzhydrol groups. Also, the polyimides containing pendent siloxane groups exhibited a reduction in thermal stability as measured by TGA as compared to the corresponding polyimides containing benzhydrol groups. The temperature of 5% weight loss by TGA for the polyimides containing pendent siloxane groups was ~380°C in air versus 490°C for the corresponding polyimides containing benzhydrol groups. All of the films in Table 1 were fingernail creasable whereas films from the corresponding benzhydrol derivative were not creasable. Polymers that contain benzhydrol units can undergo a thermally induced crosslinking reaction which may cause embrittlement.

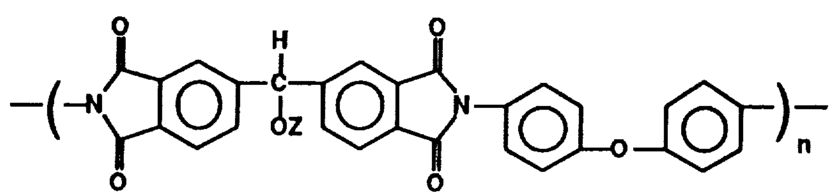
TABLE 1
POLYMER CHARACTERIZATION

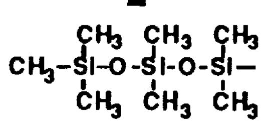


Ar	Glass Transition Temperature,°C	Film Appearance, Quality
	235	orange, translucent, tough, creasable
	219	brown, translucent, tough, creasable
	211	light tan, translucent, tough, creasable

Polymer characterization for the commercial polyimide containing benzhydrol groups and the siloxane derivative are presented in Table 2. The introduction of the pendent siloxane group in this polymer caues a similar effect, most notably the lowering of the Tg. The film of the polyimide containing pendent siloxane group was semi-translucent, but clearer than the films from the polymers in Table 1.

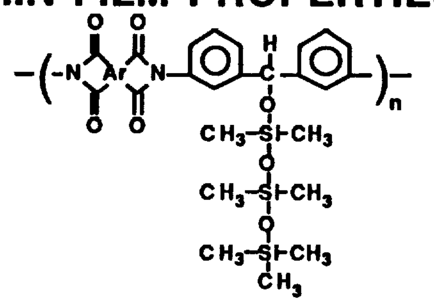
TABLE 2
POLYMER CHARACTERIZATION

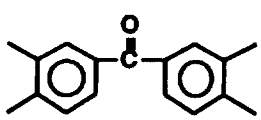
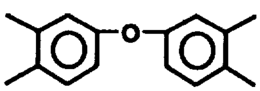
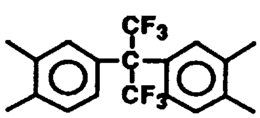


Z	Glass Transition Temperature,°C	Film Appearance, Quality
	167	yellow/green, semi-translucent, tough,creasable
H (Cemota Syntorg IP 608 Polyimide)	250	orange,clear, tough,creasable

Thin film tensile properties of the polyimides containing pendent siloxane groups are presented in Tables 3 and 4. All of the polyimides containing pendent siloxane groups exhibited a reduction in tensile strength and tensile modulus compared to the corresponding polyimides containing benzhydrol groups. However a noticeable increase in elongation to break was observed. These property changes are expected since chain to chain interactions, hydrogen bonding and the ability to crosslink are effectively eliminated by the incorporation of the pendent siloxane groups. In addition, the polyimides containing pendent siloxane groups exhibited a significantly lower dielectric constant (Table 4).

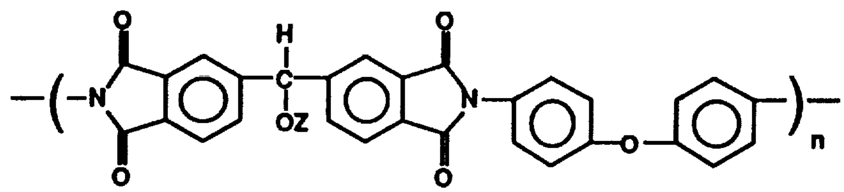
TABLE 3
THIN FILM PROPERTIES*



Ar	Tensile Strength,ksi	Tensile Modulus,ksi	Elong.,%
	12.5	390.6	9.4
	13.3	453.3	4.1
	8.2	360.6	12.0

* Tensile properties determined at 23°C.

TABLE 4
THIN FILM PROPERTIES*

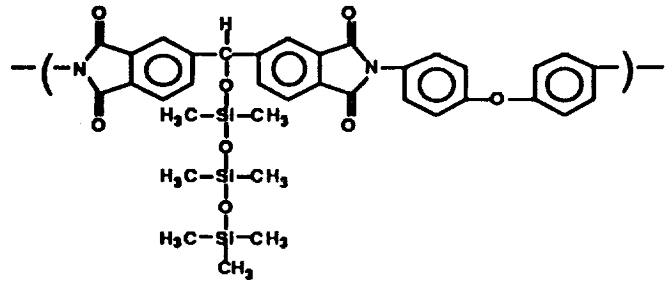


Z	Tensile Strength,ksi	Tensile Modulus,ksi	Elong.,%	Dielectric Constant
$CH_3-Si(CH_3)_2-O-Si(CH_3)_2-O-Si(CH_3)_2-$	10.9	250.0	34.3	2.8
H (Cemota Syntorg IP 608 Polyimide)	18.1	442.3	10.4	3.4

*Tensile properties determined at 23°C.

The results of the oxygen plasma exposures are given in Tables 5 and 6 for two different polyimides containing pendent siloxane groups. Each exposure was performed simultaneously with thin film samples of Kapton® and Ultem® for comparative purposes. The polyimides containing pendent siloxane groups used in this study contain ~12% silicon by weight. The polyimides containing pendent siloxane groups exhibited ~ 63-64% weight loss after 8 hours of exposure whereas Kapton® and Ultem® exhibited 100% weight loss. The weight loss data for the polyimides containing pendent siloxane groups fluctuates. For example, in Table 5 the sample after 5 hours of exposure shows a 50% weight loss and after 6 hours shows a 40% weight loss or net weight gain of ~ 10%. This weight gain is most likely due to the formation of oxides. Preliminary X-ray photoelectron spectroscopy (ESCA) results indicate the presence of silicon oxide on the film surfaces after exposure to the oxygen plasma. The polyimides containing pendent siloxane groups did not exhibit the stability (i. e., weight retention) that was anticipated. Examination of the polymer films by scanning electron microscopy (SEM) before exposure revealed the presence of pinholes. After exposure to the oxygen plasma the films were reexamined by SEM, many cracks in the films were observed all of which appeared to originate from pinholes. The pinholes act as stress concentraters and the films subsequently crack during exposure and expose fresh surface area, some of which is erroded away, resulting in additional weight loss. Pinhole-free films prepared from these polyimides containing pendent siloxane groups are expected to exhibit better resistance to the oxygen plasma.

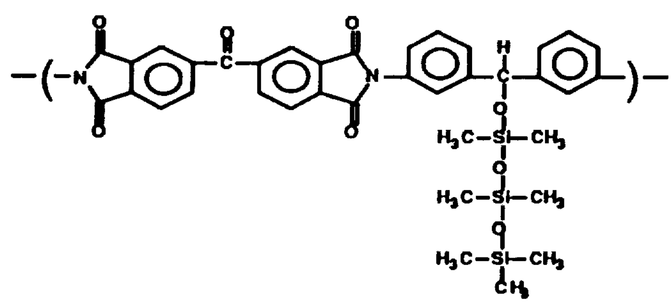
TABLE 5
ASHER EXPOSURE DATA



Weight Loss, %

Exposure Time, hr	Ultem®	Kapton®	Sample
4	37.0	38.6	6.3
5	41.2	39.8	50.5
6	62.6	43.2	40.4
8	100.0	100.0	62.2

TABLE 6
ASHER EXPOSURE DATA



Weight Loss, %

Exposure Time, hr	Ultem®	Kapton®	Sample
4	55.5	48.0	67.5
5	45.6	52.1	59.5
6	100.0	81.3	70.6
8	-----	100.0	63.8

CONCLUSIONS

A series of polyimides containing pendent siloxane groups were prepared and characterized. The T_gs, tensile moduli, thermal stability and dielectric constant of the polyimides containing pendent siloxane groups were lower than that of the corresponding precursor polyimides containing benzhydrol groups. The polyimides containing pendent siloxane groups exhibited significantly lower weight loss than Kapton® and Ultem® when exposed to an oxygen plasma. Preliminary results indicate that the polyimides containing pendent siloxane groups formed a protective silicate-type surface layer during oxygen plasma exposure.

ACKNOWLEDGMENT

The generous donation of the polyimide containing benzhydrol groups, Syntorg IP 608, by C. E. M. O. T. A., Vernaison France (I. F. P. Enterprises, Inc., New York NY) is greatfully appreciated.

REFERENCES

1). B. A. Banks, M. J. Mirtich, S. K. Rutledge and H. K. Nahra, Proc. 18th IEEE Photovoltaic Specialists Conf., (1985).
2). L. J. Leger, I. K. Spiker, J. F. Kuminecz, T. J. Ballentine and J. T. Visentine, STS Flight 5, LEO Effects Experiment, AIAA-83-2631-CP, (1983).
3). K. A. Smith, Evaluation of Oxygen Interaction with Materials (EOIM), STS-8 Atomic Oxygen Effects, AIAA-85-7021, (1985).
4). D. L. Bailey and M. Pike, U. S. Patent 2,998,406 to Union Carbide Corp., (1961).
5). V. H. Kuckertz, Die Makromolekulare Chemie, 98, 101 (1966).
6). J. K. Gilliam and H. C. Gilliam, Polymer Engineering and Science, 13(6), 447 (1973).
7). I. Yilgor, E. Yilgor, B. C. Johnson, J. Eberle, G. L. Wilkes and J. E. McGrath, Polymer Preprints, 24(2), 78 (1983).
8). S. Maudal and T. L. St. Clair, International Journal of Adhesion and Adhesives, 4(2), 87 (1984).
9). B. C. Johnson, I. Yilgor and J. E. McGrath, Polymer Preprints, 25(2), 54 (1984).
10). C. J. Lee, Society for the Advancement of Material Process and Engineering Series, 30, 52 (1985).
11). A. Berger, Ibid., 30, 64 (1985).
12). F. F. Holub, U. S. Patent 3,325,450 to The General Electric Corp., (1973).
13). J. T. Hoback and F. F. Holub, U. S. Patent 3,740,305 to The General Electric Corp., (1973).
14). A. Berger, U. S. Patent 4,011,279 to The General Electric Corp., (1977).
15). H. Sato, U. S. Patent 4,395,426 to Hitachi Chemical Company Ltd., (1982).
16). A. Berger, U. S. Patent 4,395,527 to M and T Chemicals, Inc., (1983).
17). H. Ryang, U. S. Patent 4,404,350 to The General Electric Corp., (1983).
18). Y. Nagase, S. Mori, M. Egawa and K. Matsui, Makromol. Chem. Rapid Commun., 11, 185 (1990).

N92-22463

CORROSION-PROTECTIVE COATINGS FROM ELECTRICALLY CONDUCTING POLYMERS

Karen Gebert Thompson & Coleman J. Bryan
National Aeronautics and Space Administration
Materials Science Laboratory (DM-MSL-22)
Kennedy Space Center, Florida 32899

Brian C. Benicewicz & Debra A. Wrobleski
Materials Science and Technology Division
Los Alamos National Laboratory
Los Alamos, New Mexico 87545

ABSTRACT

In a joint research effort involving the Kennedy Space Center and the Los Alamos National Laboratory, electrically conductive polymer coatings have been developed as corrosion- protective coatings for metal surfaces. At the Kennedy Space Center, the launch environment consists of marine, severe solar, and intermittent high acid/elevated temperature conditions. Electrically conductive polymer coatings have been developed which impart corrosion resistance to mild steel when exposed to saline and acidic environments. Such coatings also seem to promote corrosion resistance in areas of mild steel where scratches exist in the protective coating. Such coatings appear promising for many commercial applications.

INTRODUCTION

Research in the last decade has brought to light a new class of polymeric materials known as electrically conductive polymers. Many experts have touted this new class of materials as having the potential to combine the conductivity of a metal with the lightweight convenience and chemical resistance of a plastic. The physical and chemical properties of polymers such as high strength-to-weight ratios, toughness, low cost, molecular tailoring of desired properties, and ease of processing into films, filaments, and complex shapes make polymeric materials extremely attractive for many applications. Over the last several years, efforts to develop a new generation of stable and processable conducting polymers appear to be on the brink of success. One such research effort involving the Kennedy Space Center (KSC) and the Los Alamos National Laboratory (LANL) entails the development of corrosion-protective coatings from electrically conductive polymers. This paper discusses the development and testing of these conductive polymer coatings.

BACKGROUND

Until recently, the field of electrically conductive polymers comprised materials with virtually no processability. In the last few years, however, it was discovered that monomers based upon aniline, thiophenes, and pyrroles can be synthesized and polymerized to high molecular weight materials. Through proper control of substituents, polymers have been made that are both soluble in common organic solvents and melt processable below decomposition temperatures. Such breakthroughs demonstrate the potential of producing processable electrically conductive polymers.

The concept of using electrically active coatings for corrosion protection of metal surfaces has recently been addressed by F.C. Jain et. al. [1]. Metallic surfaces can host positive dipole layers when such surfaces adjoin appropriately doped semiconductors to form metal/semiconductor structures. These interfacial space charge layers result in an inherent electric field which opposes the flow of electrons from the metal surface to oxidizing species in the environment, thus lowering the rate of oxidation (i.e., corrosion). It is important to note that the current reduction is due to the existence of an active electronic barrier at the interface, and not to the electrical resistance of the semiconductor film. The electronic barrier may also inhibit corrosion in regions where pinholes exist in the semiconductor layer, since a finite electric field is expected to retard transfer of electrons. This same theory may apply to electrically conductive polymers used as corrosion-control coatings.

A coating with resistance to hydrochloric acid and to corrosion is needed for ground support equipment and structures at KSC. The launch environment consists of a marine, severe solar, and intermittent high acid/elevated temperature environment. The current zinc-rich coatings used on launch structures have the drawback of attack of the zinc moiety by the high concentrations of hydrochloric acid released during a Space Shuttle launch. The KSC and LANL research effort involves the synthesis of electrically conductive polymers, formulation of such polymers into coatings, and subsequent environmental and physical testing of steel specimens coated with these materials. The objective of the study is to formulate these organic coatings to provide easy application, repair, and long term resistance to the KSC launch environment.

COATING PREPARATION AND TESTING

Polymer selection

The research team has synthesized several conducting polymers and prepared solutions of suitable viscosity for casting films. Solvents, casting techniques, and drying conditions have been developed for coating steel coupons with pinhole-free films. For a material to qualify as a candidate for a corrosion-protective coating, selection criteria include ease of preparation and processing, dopability (i.e. increasing conductivity by additives serving as electron donors or acceptors), electrical conductivity, environmental stability, mechanical integrity of film, adhesion to steel, and low cost.

Several conducting polymers were synthesized during the course of the research effort. In some cases specialized monomers were synthesized before subsequent polymerization; in other cases monomers were obtained commercially. Many of the polymers considered were eliminated from the study based upon the qualification criteria listed above. For example, some of the polymers requiring specialized monomer synthesis were eliminated due to high cost to produce such materials. Recent work has encompassed development of methods for coating steel coupons with pinhole-free films of the following pi-conjugated polymers: polyaniline, poly(3-hexyl thiophene), poly(3-octyl thiophene), poly(3-thienylmethylacetate), and poly(3-thienylethylacetate). Figure 1 gives the chemical structures for these polymers.

Adhesion to steel was the main obstacle in the study. Adhesion problems were solved for many of the polymers through efforts such as investigating an appropriate blend of conductive polymer and epoxy and by applying undoped, chemically prepared polymer to the surface of the steel and subsequently doping the coated surface to the conducting state. Once steel samples were coated with candidate materials, the samples were exposed to salt water and to 0.1 M HCl. Results of such adhesion studies concentrated the work effort on polyaniline coatings, which were clearly superior to others tested.

Polyaniline

The oxidative polymerization of aniline to polyaniline was reported in the literature as early as 1862 [2]. Polyaniline was known as "aniline black" and was used as a textile dye. In fact, electrically conductive polymers are intensely colored. Researchers in France in 1967 first reported the electrical conductivity of certain members of the polyaniline family [3]. The electrical conductivity of polyaniline is a function of its oxidation and protonation states. In the polyaniline structure depicted in Figure 2, y represents reduced or benzenoid units, and 1-y represents oxidized or quinoid units. When polyaniline is composed solely of reduced units, the material is colorless and an insulator. When polyaniline is made up solely of oxidized units, the material is black in color and is a readily hydrolyzed insulator. The emeraldine form is the most conductive form and has roughly equal numbers of reduced and oxidized units.

Polyaniline was synthesized chemically according to the method reported in the literature [4]. Ammonium persulfate was the oxidizing agent used. The polyaniline powder was converted to the nonconducting emeraldine base by stirring in an ammonium hydroxide solution. The product was filtered, washed, partially dried, pulverized, and dried to a constant weight.

The emeraldine base of polyaniline can be easily dissolved in organic solvents for application to steel substrates. Once polyaniline is doped to the conducting state, however, the material has limited solubility in organic solvents. Consequently, the surface of the mild steel samples were initially coated with the undoped, chemically prepared polyaniline. This coating of the mild steel coupons was accomplished using solutions of polyaniline in 1-methyl-2-pyrrolidinone (NMP). Dip-coating methods were first used for applying the polyaniline/NMP solutions to the steel, followed by development of a spray method. Both methods provided good coverage and good adhesion of the polyaniline to the steel. Upon drying, the coatings were 1-2 mils (0.001-0.002 in.; 0.03-0.05 mm.) thick.

Once dry, the undoped polyaniline coating was doped to the conducting state. By increasing the electrical conductivity of the polymer, dopants provide the proper electronic environment to impart corrosion resistance and acid resistance to the film. Over twenty-five different dopants were evaluated during the study. The dopants which gave the best results were tetracyanoethylene (TCNE), zinc nitrate, and p-toluenesulfonic acid.

After the coating was doped to a conducting state, a topcoat of cross-linked epoxy was applied to the samples in order to impart improved abrasion resistance to the coating. The epoxy topcoat used on the samples discussed in this paper was Ciba-Geigy Bisphenol A GY 2600 resin cured with a cycloaliphatic/aliphatic amine hardener XU265. The resultant coating was designed to provide the proper electronic environment as well as coating toughness and resistance to harsh environmental conditions.

Corrosion Testing

Polymer-coated steel coupons were tested for corrosion resistance in two different environments via gas/liquid cells. One environment consisted of placing each coupon in an individual vial containing enough 3.5% NaCl solution to cover the coated portion of the coupon. All vials were capped with a rubber septum into which air was bubbled to ensure oxygenation of the solution. In the second environment, a 0.1 M hydrochloric acid (HCl) solution was used in place of a saline solution. Photographs were taken of the samples before exposure to the above environments as well as throughout the testing. In some cases, the tests were carried out for 12 weeks.

To establish a baseline of the corrosion resistance of the conductive polymer-coated samples, control samples of mild steel coated solely with the epoxy coating were tested. The Ciba Geigy epoxy material was chosen because of its use in power plants for coating interior surfaces of stacks emitting sulfur dioxide. The material is reported to exhibit acid resistance [5]. The control samples were tested with the conductive polymer samples in the corrosion tests.

Figure 3 shows photographs of samples before and after twelve weeks of exposure to aerated 3.5% NaCl solution. The photograph on the top left shows mild steel coated only with epoxy before exposure. The sample shown on the top right was sprayed with undoped polyaniline in NMP, then doped with p- toluenesulfonic acid, followed by application of an epoxy topcoat. The sample shown on the bottom left is an epoxy control sample after twelve weeks exposure to the saline solution. Corrosion was evident with the control sample, with pitting throughout the sample and mass loss from the edges of the steel sample observed. The sample on the bottom right is a polyaniline/epoxy sample. No evidence of corrosion was seen, with the edges of the sample still intact and showing no mass loss. Since the polyaniline is dark in color, the coating was scraped off to verify that no corrosion was present. In fact, the polyaniline coating adhered to the substrate so tenaciously that it was difficult to scrape the coating from the steel.

Figure 4 shows photographs of samples before and after twelve weeks of exposure to aerated 0.1 M HCl solution. The photograph on the top left shows mild steel coated only with epoxy before exposure. The photograph on the top right shows mild steel which was sprayed with undoped polyaniline in NMP, then doped with tetracyanoethylene (TCNE), followed by application of an epoxy topcoat. Both samples were scribed before exposure to acid. The photograph on the bottom left shows extensive corrosion on the epoxy control sample after eight weeks in acid. The photograph on the bottom right shows the polyaniline/epoxy sample. No evidence of corrosion was seen, with the scratched surface still shiny.

Many environmental tests have been carried out in saline and acidic oxidizing environments. A marked improvement in the corrosion resistance of mild steel has been observed when using the electrically conductive polymer coatings developed in this research program as compared to mild steel coated solely with epoxy.

Additional Environmental and Physical Testing

Candidate coatings were initially screened by the corrosion tests described above using aerated saline and hydrochloric acid solutions. The best candidate materials are undergoing further testing for determining effective corrosion resistance for mild steel. Testing methods include ultraviolet (UV) radiation testing, electrochemical corrosion testing, accelerated corrosion testing in a salt fog chamber, long term exposure at the KSC beach corrosion testing site, pitting corrosion tests in ferric chloride solution, and electrochemical impedance spectroscopy.

Samples are subjected to high intensity UV radiation in an Atlas Electric Devices Weatherometer. Such UV exposure is used to determine outdoor weathering properties of a material in regard to sunlight and rain. Specimens are exposed to 0.35 Watt/m^2 of 340 nanometer wavelength UV radiation.

The electrochemical corrosion testing utilizes a Model 351-2 Corrosion Measurement System, manufactured by EG&G Princeton Applied Research. The electrochemical cell includes a saturated calomel reference electrode, two graphite rod counter electrodes, a metal specimen working electrode, and a bubbler/vent tube. The electrolyte consists of varied concentrations of HCl solution plus 3.55% NaCl. Electrochemical tests performed include determination of corrosion potential, polarization resistance (per test procedure in ASTM G59 [6],) and cyclic polarization (per test procedure in ASTM G61 [7]).

Accelerated corrosion testing is carried out in an Atlas Corrosive Fog Exposure System Model SF-2000, manufactured by Atlas Electric Devices Company. The solution used for salt fog exposure is a standard 5% NaCl mixture. After each week of exposure to salt fog, the specimens are immersed for one minute in a 1.0 M HCl/alumina mixture to simulate the solid rocket booster effluent created during launch of the Space Shuttle. Following immersion in HCl/alumina, the specimens are allowed to drain and dry overnight and then are returned to the salt fog chamber for the next one week cycle. The inspection procedure includes cleaning, weighing, and visual characterization of the corrosion occurring.

All beach exposure testing is carried out at the KSC Beach Corrosion Test Site, which is located on the Atlantic Ocean approximately one mile south of Launch Complex 39A at KSC. The test site is approximately 100 feet from the mean high tide line, with the orientation of the samples facing east toward the ocean at a 45 degree angle, to receive the full extent of sun, rain, and sea spray. The beach exposure test procedure is based on the test method described by ASTM G50 [8], with the addition of an acid spray. Every two weeks the specimens are sprayed with a 1.0 M HCl/alumina powder slurry, which thoroughly wets the surface of the specimen and which is allowed to remain on the surface of the specimen until dry or rinsed off by rain. The inspection procedure includes cleaning, weighing, and visual characterization of the corrosion.

The pitting corrosion testing in ferric chloride solution is based on ASTM G48, Method A [9]. Specimens are immersed in an aqueous solution of ferric chloride for 72 hours. Following a water rinse and removal of corrosion products, the specimens are dipped in acetone or alcohol and allowed to air dry. Each specimen is weighed and examined visually and at low magnification for signs of pitting.

Electrochemical evaluation of the conductive polymer/steel interface is performed, using alternating current (ac) and direct current (dc) methods. Such methods are used as accelerated test methods. AC impedance measurements made over a range of frequencies utilizes a Model 378 electrochemical measurement system manufactured by EG&G Princeton Applied Research. The dc work utilizes a Model 1000 system manufactured by EG&G. Samples are immersed in aerated, natural seawater collected from the Atlantic Ocean at Cape Canveral, Florida. Electrochemical ac impedance measurements are made at approximately one week intervals for six weeks. Direct current linear polarization resistance measurements are also made periodically, and the corrosion potential is measured as well. Values for polarization resistance are obtained from Nyquist diagrams and from the dc linear polarization data. Values are also calculated for the coating capacitance.

FUTURE STUDIES

Current work is producing improved coatings with greater environmental stability. This work will continue, and corrosion resistance and physical properties of the coatings will be evaluated. Efforts are also underway to establish a working agreement to aid in the commercialization of these coatings. The aim of this agreement would be to develop the capability of manufacturing large batches of these coatings and to subsequently develop methods for coating large structures.

COMMERCIAL APPLICATIONS

Corrosion-protective coatings from electrically-conducting polymers have many commercial applications. Most applications considered thus far regard corrosion protection to equipment exposed to heat, sunlight, saline environments, and other outdoor exposure concerns. There is a great need for such coatings for bridges, for example. Another use is for coating rebar used in concrete. Improved coatings for underground storage tanks are greatly needed. The automotive industry also needs improved coatings--especially for customers living in sea coast environments. The authors of this paper were recently contacted by industry regarding the need for a protective coating on oceanic drilling platform equipment. A need which deviates from outdoor environmental concerns involves an application in a new electrically erodable printing process which requires conductive paths in specific locations on the printing plate. Electrically conductive polymer coatings afford advantages over existing materials in properties including bonding ability, processability, durability, and greater strength-to-mass ratio. In addition to the suggested applications listed herein, there are many commercial applications for an easily processable, electrically-conductive polymer coating with good adhesion and good environmental resistance properties.

REFERENCES

1. Jain, F.C.; Rosato, J.J.; Kalonia, K.S.; Agarwala, V.S.; Corrosion 1986, 42, 700.

2. Letheby, H.; J. Chem. Soc. 15, 161 (1862).

3. Jozefowicz, M.; Yu, L.T.; Belorgey, G.; Buvet, R.; J. of Polymer Science. Part C 16, 2943 (1967).

4. Cao, Y.; Andreatta, A.; Heeger, A.; Smith, P.; Polymer, 30 (12) 2305-11 (1989).

5. "Organic Coatings in Simulated Flue Gas Desulfurization Environments", Electric Power Research Institute Report #CS-5449, Research Project 1871-5, October 1987, H. Leidheiser, Fr., M.L. White, D.J. Mills.

6. ASTM G59-78, Standard Practice for Conducting Potentiodynamic Polarization Resistance Measurements, 1986 Annual Book of ASTM Standards, Volume 03.02, ASTM, Philadelphia, PA, 1986.

7. ASTM G61-78, Standard Practice For Conducting Cyclic Potentiodynamic Polarization Measurements For Localized Corrosion, 1986 Annual Book of ASTM Standards, Volume 03.02, ASTM, Philadelphia, PA, 1986.

8. ASTM G50-76, Standard Practice for Conducting Atmospheric Corrosion Tests on Metals, 1986 Annual Book of ASTM Standards, Volume 03.02, ASTM, Philadelphia, PA, 1986.

9. ASTM G48-76, Standard Test Methods For Pitting and Crevice Corrosion Resistance of Stainless Steels and Related Alloys By The Use Of Ferric Chloride Solution, 1986 Annual Book of ASTM Standards, Volume 03.02, ASTM, Philadelphia, PA, 1986.

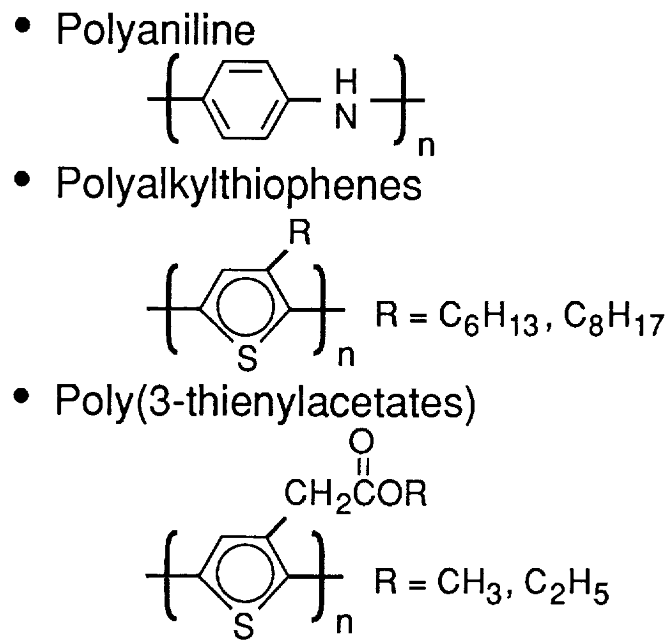


FIGURE 1. ELECTRICALLY CONDUCTING POLYMERS STUDIED

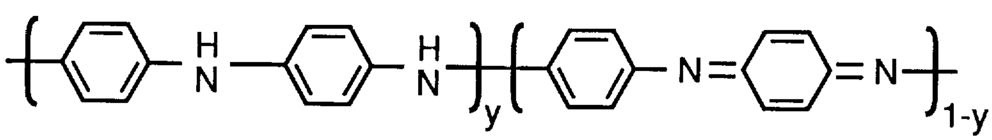
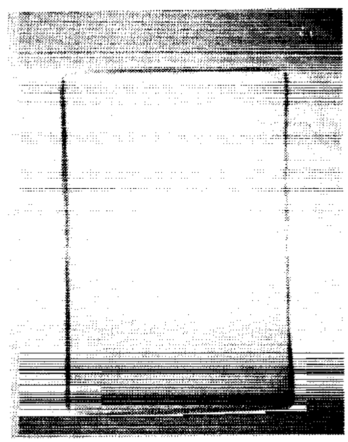
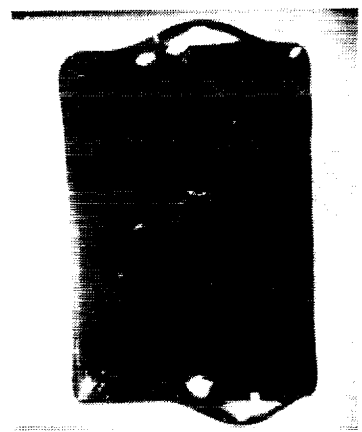
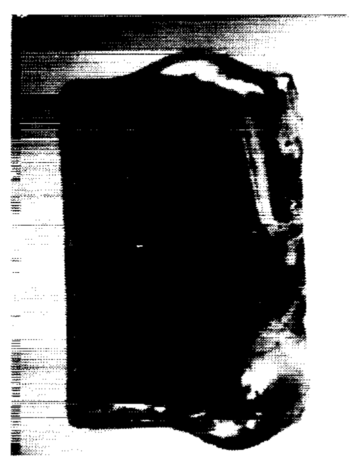
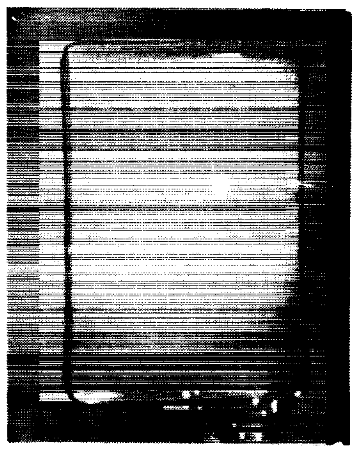
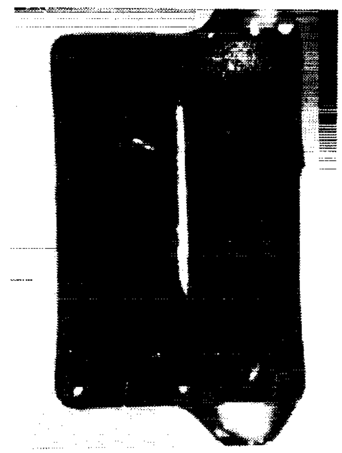
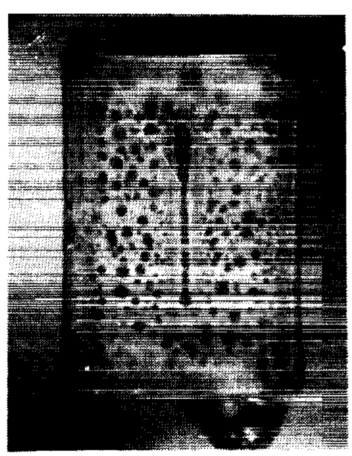
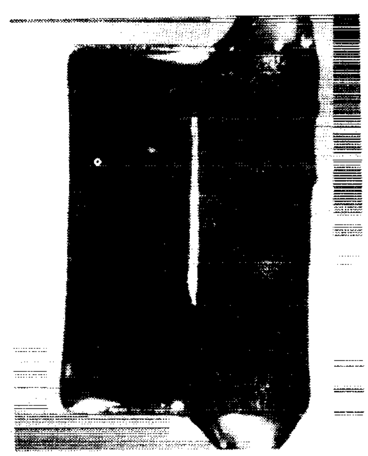


FIGURE 2. POLYANILINE STRUCTURE IN UNDOPED FORM, SHOWING REDUCED AND OXIDIZED UNITS

	Mild Steel Coated Solely With Epoxy	Mild Steel With Polyaniline, p-$TolSO_3H$* Dopant, & Epoxy Topcoat
Photographs of Specimens Before Corrosion Testing		
Photographs of Specimens After 12 Weeks Corrosion Testing		

* p-$TolSO_3H$ = p-Toluenesulfonic acid

FIGURE 3. CORROSION TESTING OF COATED STEEL SPECIMENS IN AERATED 3.5 % NaCl. (Magnification: 3X)

	Mild Steel Coated Solely With Epoxy	Mild Steel With Polyaniline, TCNE* Dopant, & Epoxy Topcoat
Photographs of Specimens Before Corrosion Testing		
Photographs of Specimens After 8 Weeks Corrosion Testing		

* TCNE = Tetracyanoethylene

FIGURE 4. CORROSION TESTING OF SCRIBED SPECIMENS IN AERATED 0.1 M HCl. (Magnification: 3X)

MEDICAL ADVANCES: COMPUTERS IN MEDICINE

(Session B5/Room C4)

Wednesday December 4, 1991

- Computation of Incompressible Viscous Flows Through Artificial Heart Devices
- Computer Interfaces for the Visually Impaired
- Extended Attention Span Training System
- Man/Machine Interaction Dynamics and Performance Analysis Capability

N92-22464

COMPUTATION OF INCOMPRESSIBLE VISCOUS FLOWS THROUGH ARTIFICIAL HEART DEVICES WITH MOVING BOUNDARIES

Cetin Kiris
MCAT Institute
NASA Ames Research Center, Moffett Field, CA

Stuart Rogers, Dochan Kwak
NASA Ames Research Center, Moffett Field, CA

I-Dee Chang
Stanford University, Stanford, CA

ABSTRACT

The current work illustrates the extension of computational fluid dynamics techniques to artificial heart flow simulation. Unsteady incompressible Navier-Stokes equations written in three-dimensional generalized curvilinear coordinates are solved iteratively at each physical time step until the incompressibility condition is satisfied. The solution method is based on the pseudo-compressibility approach and uses an implicit-upwind differencing scheme together with the Gauss-Seidel line relaxation method. The efficiency and robustness of the time-accurate formulation of the numerical algorithm are tested by computing the flow through model geometries. A channel flow with a moving indentation is computed and validated with experimental measurements and other numerical solutions. In order to handle the geometric complexity and the moving boundary problems, a zonal method and an overlapped grid embedding scheme are employed, respectively. Steady-state solutions for the flow through a tilting-disk heart valve are compared against experimental measurements. Good agreement is obtained. The flow computation during the valve opening and closing is carried out to illustrate the moving boundary capability. Aided by experimental evidence, the flow through an entire Penn State artificial heart model is computed.

I. INTRODUCTION

With the advent of supercomputer hardware as well as fast numerical methods, researchers in the field of computational fluid dynamics (CFD) tackle more complicated problems than ever before. With these new capabilities, CFD has become an essential part of aerospace research and design. For example, the incompressible flow solver developed by Kwak et al [1] was extensively used for simulating the flow through space shuttle main engine power head components. The redesign of the space shuttle main engine hot gas manifold, guided by the computations of Chang et al. [2] illustrates the usefullness of CFD in the aerospace research.

Extending the CFD technology developed for the aerospace industry to artificial heart simulations will open a new route in the artificial heart research. Artificial heart devices have been used widely since the early 1960s to replace or to assist natural organs. The replacement can be a heart valve or a total artificial heart. However, the prosthetic devices are found to be less efficient than natural organs and various problems have been found during clinical observations. The most serious problems are believed to be directly associated with the flow fields created inside the artificial heart devices. Major difficulties originating from fluid dynamics phenomena include : 1)- Separated and secondary flow regions cause clotting; 2)- High turbulent shear stress can damage the red blood cells; 3)- Large pressure losses across the valves prevent the heart from working efficiently. Several experimental studies[3-5] on commonly used valve geometries have pointed out the adverse effects of the stagnation and recirculation regions on blood flow. Although the experimental studies played an important role in the design process of these devices, they can provide flow characteristics for only limited regions of the flow field. In addition, the experimental measurements are very difficult because of the moving boundaries in the artificial heart devices. Having detailed knowledge of the flow quantities can help a design engineer improve the artificial heart and valve geometries, where a smooth flow is desired. The development of such a numerical simulation tool, which can be used in the artificial heart research programs, is initiated in the present study.

 PRECEDING PAGE BLANK NOT FILMED

Computational studies of blood flow in hearts and heart valves have been quite limited. The most notable work has been performed by Peskin and his co-workers.[6-8] Their main effort has been directed to simulations of the natural heart and heart valves combining Eulerian flow equations and a Lagrangian description of heart walls and valves. Peskin and McQueen[6] modeled the prosthetic heart valves in the numerical simulation of the flow in the natural heart. They used boundary forces derived from the energy function in order to model valve opening and closing, and they also modeled the elastic behavior of the walls. Their solutions were obtained for low Reynolds numbers in two dimensions using a square cartesian mesh. McCracken and Peskin[7] applied a combined vortex-grid method for the blood flow through the mitral valve in two dimensions. Peskin and McQueen[8] demonstrated the capability of modeling the elastic behavior of heart muscles by applying their extended three dimensional solution procedure to a toroidal tube. Although determining the elastic behavior of the walls is an important task in the natural heart study, the boundary motion in many artificial heart devices, such as the Penn State electric artificial heart, can be well defined.[9] Therefore, the present study is focused on the fluid problem with prescribed body motion.

Underwood and Mueller[10] obtained the flow characteristics for the Kay-Shiley disk type valve using the stream function-vorticity formulation. Their results showed agreement with experimental data up to a Reynolds number of 600. Idelsohn, Costa, and Ponso[11] modeled the flow through the Kay-Shiley caged disk, Starr-Edwars caged ball, and Bjork-Shiley tilting disk valves and compared their performance. Turbulent flow through trileaflet aortic heart valves was simulated by Stevensen et al [12]. Most numerical studies assumed that the flow through the heart valve was two-dimensional. Additionally, the valve opening and closing motion was neglected; only the flow through a fixed valve position was studied. In reality, the geometry is three-dimensional, and the flow through heart valves involves moving boundaries.

The current study proposes the development of a computational procedure simulating steady and unsteady, three-dimensional flows through artificial hearts and heart valves with moving boundaries. In the next sections, the method of solution is summarized followed by demonstration how this CFD procedure can be used for probing the flow through artificial heart devices.

II. METHOD OF SOLUTION

The flow through artificial heart devices is unsteady, viscous, and incompressible. In the present study, the non-Newtonian nature of the blood is neglected and the flow is described by the three-dimensional incompressible Navier-Stokes equations. Numerical simulation of such flows is a very challenging problem in computational fluid dynamics. In addition to the geometric complexity, obtaining time-dependent solutions of the incompressible Navier-Stokes equations with moving boundaries poses many difficult numerical problems.

Because the pressure and velocity fields are not directly coupled due to the lack of a pressure term in the continuity equation, numerical solution of the incompressible Navier-Stokes equations requires special attention in order to satisfy the divergence-free constraint on the velocity field. The most widely used methods which use primative variables are fractional-step and pseudocompressibility techniques. In fractional step method, the auxiliary velocity field is solved by using the momentum equations. Then, a Poisson equation for pressure is formed by taking the divergence of the momentum equations and by using a divergence-free velocity field constraint. Solving the Poisson equation for pressure efficiently in three-dimensional curvilinear coordinates is the most important feature of the fractional step method.[13] One way to avoid the numerical difficulty originated by the elliptic nature of the problem is to use a pseudocompressibility method. With the pseudocompressibility method, the elliptic-parabolic type equations are transformed into hyperbolic-parabolic type equations. Well established solution algorithms developed for compressible flows can be utilized to solve the resulting equations.

In the present study, the pseudocompressibility approach is used and the time accuracy is attained by iterating in pseudo-time until the divergence of velocity is driven toward zero to within a specified error tolerance. Here, the time derivatives in the momentum equations are differenced using a second-order, three-point, backward-difference formula. The numerical method uses a second-order central difference for viscous terms and a higher order flux-difference splitting for the convective terms. The resulting matrix equation is solved iteratively by using a nonfactored line relaxation scheme, which maintains stability and allows a large pseudo-time step to be taken.

At each sweep direction, a tridiagonal matrix is formed and off line terms of the matrix equation are moved to the right-hand side of the equation. Details of the governing equations and numerical method are given in Refs. 14-15.

One of the biggest difficulties in the simulation of flows in complicated three- dimensional configurations is the discretization of the physical domain. The problem becomes more severe if one body in the domain of interest moves relative to another one as is seen in the tilting disk valve and the Penn State artificial heart geometries. The use of a zonal approach is a practical solution if the grids are stationary. For more general applications, a Chimera grid embedding technique[16] provides a greater flexibility for the grid motion. This technique is employed for flow computations through a tilting disk valve and the Penn State Artificial heart model.

III. COMPUTED RESULTS

One of the goals of this study is to simulate the flow through a realistic model of an artificial heart. Since the geometry and the flow physics are complicated, the computational procedure is validated by solving several simpler problems which characterize the flow in various parts of an artificial heart. As a first step, an idealized 2-D pump model was chosen to demonstrate the capability of the time-accurate formulation under a moving grid condition. Geometry of this model and computed results are presented in Ref. 14. Channel flow with an indented wall, the flow through a tilting disk heart valve and the flow through the Penn State artificial heart model are included in this section.

Channel Flow with a Moving Indentation

Channel flow with an asymmetric oscillating indentation was experimentally studied by Pedley and Stephanoff,[17] and was numerically simulated by Ralph and Pedley.[18] In the experiment, the channel wall was rigid everywhere except for the indentation which is made of a thick rubber membrane. The experiment shows that flow to be two-dimensional near the midplane and so the computation is done in two dimensions.

Figure 1 illustrates the instantaneous streamlines plotted at several nondimensional times for $Re = 600$ and $St = 0.057$. The Reynolds number is based on the channel height, a, and the average velocity at the entrance of the channel, U. Strouhal number is defined as $St = af/U$ where f is the oscillation frequency. At the beginning of the cycle the flow downstream of the indentation is parallel to the channel walls. A single eddy is formed at the sloping wall of the indentation during the first half of the cycle. The streamlines at the core flow are lifted slightly upwards as shown in Fig. l-a. This is the beginning of the wavy flow patterns of the core flow. The formation of a second separated eddy on the opposite wall can be seell in Fig. l-b. In later stages, the double row of eddies along the lower and upper wall of the channel is observed. At the end of the cycle the vortices are swept downstream (Fig. l-g), and the residual vortices are not strong enough to affect the next cycle. In fact, the flow patterns of the first and second cycles are quite similar. Consequently, the flow is assumed periodic in time, even at the first cycle. Another interesting phenomenon observed in the present study as well as observed in the experimental and other numerical studies is the eddy- doubling which can be seen in Figs. l-c through l-e. It occurred in the second, third, and fourth vortices from the indentation. In eddy-doubling, a single eddy splits into two co-rotating eddies.

The time evolution of the vortices' centers is compared against the experimental and other numerical findings. The first four vortices are called Vortices A, B, C, and D, and are shown in Fig. 2-b. The distance between the indentation wall and the center of these vortices is measured from the instantaneous streamlines and plotted versus time in Fig. 2-a. The dashed lines represent the present computations. The solid lines denote computational results from the fractional step approach implemented by Rosenfeld.[13] Dotted lines are numerical results of stream function vorticity formulations from Ralph and Pedley.[18] Experimental measurements by Pedley and Stephanoff [17] are represented by square symbols. The agreement between numerical and experimental results is fairly good. There is a discrepancy between numerical results and experimental measurements about the location of the vortex A. The present results and Rosenfeld's computations predict the location of vortex A 0.4 units closer to the indentation than experimental findings. However, the locations of the remaining vortices are correctly predicted if the distance is measured from the center of the vortex A. This underprediction of the separation length of vortex A is thought to be caused by an inaccurate description of the indentation wall shape. The present study and Rosenfeld's study used the same grid and the same wall shape. Even though the solution algorithms of the two computational studies are completely different, the agreement between the numerical results is good. For these reasons, only the location of vortices B, C, and D are compared with the experimental measurements.

Flow Through Tilting Disk Valve

This problem was chosen to develop and validate a procedure which will be used for the valve region of the Penn State artificial heart. In the Bjork-Shiley tilting disk heart valve, the tilting disk is placed in front of the sinus region of the human aorta. The aortic root has three sinuses about 120 degrees apart from one another. The tilting disk valve model used in this computation is simplified by assuming that the sinus region of the aorta has a circular cross-section. The cage and struts which hold the free-floating disk inside of the sewing ring are not included in the geometry. It is also assumed that the walls do not have an elastic deformation. The channel length is taken to be five aorta diameters long. The computational geometry used in these unsteady flow computations is given in Fig. 3. The disk motion is illustrated by showing three different positions of the disk, at angles of 75, 50, and 30 degrees as measured from the centerline of the aorta. The tilting disk is allowed to rotate about the horizontal axis that is 1/6 of a disk diameter below the center of the disk. Because of this asymmetric disk orientation, the flow is three dimensional.

The Chimera grid embedding technique, which has been successfully used for external flow problems, has been employed by using two overlapped grids as shown in Fig. 4. Grid 1 occupies the whole region in the aorta from entrance to exit, and remains stationary. Grid 2 wraps around the tilting disk, and moves with the disk. In the Chimera grid embedding technique, grid points which lie within the disk geometry and outside the channel grid are excluded from the solution process. These excluded points are called hole points, and the immediate neighbors of the hole points are called fringe points. The information is passed from one grid to another one via fringe and grid boundary points by interpolating the dependent variables. Tri-linear interpolation is used in the present computations. In order to distinguish the hole and fringe points from regular computational points, an IBLANK array is used in the flow solver. For hole, grid boundary, and fringe points IBLANK is set to zero, otherwise it is set to one. In order to exclude the hole and grid boundary points from the solution procedure, the coefficients of the system of algebraic equations and right-hand-side terms are multiplied by the IBLANK value. If the grid point is a hole, an outer boundary, or a fringe point the value of (1—IBLANK) is added to the main diagonal of the matrix equation.

Presented here are the results of steady flow with a fixed disk angle and unsteady flow with the disk motion in the configuration described above. The problems are nondimensionalized by using the entrance diameter as a unit length, and the average inflow velocity as the unit velocity. In order to reduce the computational effort and memory size, the inflow and outflow boundaries are placed a short distance from the region of interest in comparison with the boundaries in the experimental studies. In addition, the exact shape of the sinus region of the aorta used in the experiments is not known. These discrepancies could lead to slight differences between present computations and experimental measurements.

Steady-state calculations for the 30 degree disk orientation have been carried out for Reynolds numbers in the range of 2000 to 6000, in which experimental data are available. The Reynolds number is based on the diameter and the mean velocity at the entrance of the channel. A mixing length algebraic turbulence model which is used for incompressible flow through the space shuttle main engine turnaround duct[2] is utilized. Figure 5 shows the pressure drop across the Bjork-Shiley tilting disk valve at different flow rates of physiological interest. The computed and measured axial velocity profiles at 42 mm downstream from the disk are shown in Fig. 6. Axial velocity profiles are plotted in the horizontal plane through the center of the channel. The numerical results are shown with dots and the experimental results are shown with triangles. The numerical results compare favorably with the experimental measurements.[3] The largest discrepancy is seen near the walls, where the boundary layer is overestimated by the calculation. Figure 7 shows the velocity vectors at five longitudinal stations. The flow, which is directed to the upper part of the aorta, generates vortices in the sinus region of the aorta and a large separated region along the lower wall of the aorta. Since separated and low flow regions have potential for thrombus formation, clotting may occur on the upper sinus region and the lower wall of the aorta. Figliola and Mueller[5] also present mean velocity profiles, which show similar flow characteristics to those indicated by by computational study, at several locations. They computed the shear stress from the measured velocity field and observed that the maximum shear occurs at the top wall downstream of the sinus region of the aorta. This is in agreement with the velocity plot shown in Fig. 7, in which there are large velocity components just off the wall in that location. Particle traces in Fig. 8 indicate that the flow does not separate adjacent to the tilting disk;. The tilting disk separates the flow into a major flow region, which is along the upper wall of the tube, and a minor flow region along the lower wall of the tube. Separation, reverse flow and swirling motion mostly occur in the minor flow region.

Figure 9 shows vorticity magnitude contours on the surface of the tube, outflow surface, and inflow surface of the disk, respectively. It is assumed that maximum vorticity magnitudes indicate the regions of high shear. The sewing ring surface and the edges of the disk are the regions having maximum vorticity magnitude. The upper wall of the channel also has considerably high vorticity magnitudes.

Unsteady flow calculations have been carried out in order to demonstrate and analyze the flow during disk opening and closing. For the present computation, one cycle of valve opening and closing requires 70 physical time steps. During each time step, subiterations are carried out until the maximum divergence of velocity and maximum residual drop below $10.^{-3}$ The computing time required for one cycle of the valve opening and closing is approximately 5 Cray-YMP hours. During the valve opening, inflow velocity is imposed at the entrance of the channel. The inflow velocity is chosen as a sine function in time. The disk rotation is specified as a linear function in time. Since the forces acting on the disk are known from the numerical solution, the disk rotation angle can be determined. However, there is a limitation in the disk rotation angle. For large disk rotation angle, some information may be lost between the grids when the grid embedding technique is used. In order to prevent the information loss, the maximum allowed disk rotation angle at each physical time step is taken to be less than three degrees.

Figures lO-a through lO-c illustrate the velocity vectors on the lateral symmetry plane at t/T = 0.128, 0.257, and 0.385 respectively. The velocities are very high in the region between the disk and the channel wall as shown in Fig. lO-a. During the disk opening, two vortices are formed at the upper and lower edges of the disk. The flow starts to separate behind the disk and reattaches to the wall as shown in Fig. lO-b. The stagnation region behind the disk moves downstream as the disk rotates. Highly skewed velocity profiles are seen downstream from the disk as illustrated in Fig. lO-c. The growth of the vortices has also been observed in the sinus region of the aorta while the flow opens the valve. Along the lower wall a separation region is formed.

Artificial Heart Flow

The geometry of the Penn State artificial heart model is composed of a cylindrical chamber with two tube extensions (see Fig. 11). The inflow (mitral) and outflow (aortic) tubes contain concave tilting disks which open and close to act as valves. In the computational model, tilting disk mitral valve orientation in time was obtained from the experimental data provided by the Penn State University. The aortic valve orientation in time was approximated to mitral valve orientation with a phase difference. The pumping action is provided by a pusher plate whose velocity is sinusoidal in time. Pusher plate diameter is 7.26 cm, with a stroke length of 2.28 cm. The problem is nondimensionalized with the inflow tube diameter, which is 2.54 cm, and a unit velocity of 20 cm/sec. In the computational study, the Reynolds number based on the unit length and velocity is 900. Initially, the flow was started at rest, and four cycles of the pumping action were completed using a Cray-YMP computer at NASA-Ames Research Center. One cycle of the pusher plate's motion required 240 physical time steps. At each time step, the equations were iterated until the maximum divergence of velocity was reduced below $10.^{-2}$ During most of the cycle 10-20 subiterations were required (for more detail, see Refs. 14-15).

In order to handle the geometric complexity and the moving boundary problems, a zonal method and an overlapped grid embedding[16] scheme are employed, respectively. In the zonal method, a complex computational domain is divided into several simple subdomains. The overlapped grid embedding scheme allows subdomains to move relative to each other, and provides great flexibility when the boundary movement creates large displacements. The computational grid for this heart model is shown in Fig. 11. Grid 1 is generated for the pusher plate and moves with it. Grid 2 occupies the chamber and remains stationary. Grid 3 and Grid 5 are for the inflow and outflow tube extensions, respectively. Grid points for the tubes and grid points for the chamber are overlapped on three common planes. In other words, the grid points for the tubes start three stencils inside of the chamber outer boundaries. Zonal boundary conditions are used at the interface boundaries. Grid 4 and Grid 6 wrap around tilting disks, and move with the disks. An overlapped grid embedding scheme is employed between moving grids and stationary grids.

Computed results presented next are mainly to demonstrate how CFD can be used to understand the flow in the artificial heart, and comparison with experiment is qualitative. Unsteady particle traces are illustrated in Fig. 12. The particles are released near the inflow valve at the beginning of the fourth cycle. The figure is plotted at non-dimensional time t/T = 0.45 into the period at which time the pusher plate is close to its lowest position, where T denotes the period for the pusher plate's motion. Figure 12 shows that the flow creates a strong vortex in the center region of the chamber. The particles have a swirling motion against the back wall opposite the mitral valve opening. The flow also separates at the connection region of the chamber and the inflow tube. Figure 13 shows the computed velocity vectors on the horizontal mid-plane at non-dimensional time 0.375. At that time the pusher plate is moving down, and the mitral valve is opened. Figure 13 also indicates the presence of strong circulation in the chamber. However, the three dimensional structure of the flow can not be seen clearly because the vectors are plotted on a two dimensional plane.

The strong vortex in the center of the chamber is actually created where the chamber and inflow tube are connected. The vortex moves to the core of the chamber in time. Experimental measurements by Baldwin and Tarbel[19] are illustrated in Fig. 14. Since the computational study does not include the blood sac inside the chamber, the comparison between experimental and computational results is qualitative. In addition, the Reynolds number in the experimental study is 1.7 times larger than the Reynolds number in the computational study because the flow is assumed laminar in the present computations. The biggest discrepancy between experimental measurements and computational results is the location of the vortex core in the chamber. In Fig. 13, the vortex is off center in the chamber. In Fig. 14, the vortex is located almost in the center of the chamber. Another difference can be seen in the wake region of the mitral valve. Since the Reynolds number in the computational study is lower, the wake is not as strong as the wake in Fig. 14. The Reynolds number in the future computational study will be increased by including the turbulence modeling in order to have a quantitative comparison between experimental and computational results.

During the second half of the cycle time, the pusher plate moves upward, and the outflow valve is opened. A top view of the computed velocity vectors on the horizontal plane at t/T = 0.625 is plotted in Fig. 15. Since the inflow valve is closed, residual eddies are quite large near the disk. However, they are quickly weakened as the pressure builds up inside the chamber. Measured velocity vectors at t/T = 0.625 are shown in Fig. 16.

Figure 17 shows additional vortex structures in the vertical plane of the chamber at non-dimensional time 0.25. That vortex structure causes the swirling motion of the particles previously mentioned in Fig. 12. The swirling flow in the outflow tube is shown in Fig. 18. The velocity vectors in the cross-sectional plane of the outflow tube at t/T = 0.75 are plotted. The cross-sectional plane is one non-dimensional unit downstream of the tilting disk valve, and the normal vector of that cross-section is in positive x-direction shown in Fig. 13.

SUMMARY

An efficient and robust solution procedure is developed, and validated for numerical simulations of internal flows through artificial heart devices. The solution procedure for unsteady incompressible viscous flow computations has been extended with the incorporation of the grid embedding approach. This has been used to simulate the flow through a tilting disk heart valve and the flow through the Penn State artificial heart model. Separated and secondary flow regions have been pointed out in the tilting disk heart valve and artificial heart flow simulations.

The vortex created in the central portion of the Penn State artificial heart provides good wall washing over the entire chamber. The present capability of simulating complicated internal flow problems with moving boundaries is demonstrated. The procedure developed in this study is quite general and applicable for various types of artificial heart and valve geometries. It is hoped that researchers and designers in the artificial heart research may further benefit from the computational ability obtained in the current work.

ACKNOWLEDGMENTS

This work is partially supported by the NASA Technology Utilization office.

REFERENCES

[1] Kwak, D., Chang, J. L C., Shanks, S. P., and Chakravarthy, S., "A Three- Dimensional Incompressible Navier-Stokes Flow Solver Using Primitive Variables," Incompressible Navier-Stokes Equations ", AIAA *Journal*, Vol 24, no. 3, pp. 390-396, 1977.

[2] Chang, J. L C.,Kwak, D., Rogers, S. E., and Yang, R.-J., " Numerical Simulation Methods of Incompressible Flows and an Application to the Space Shuttle Main Engine", *Int. J. Num. Meth. in Fluids,* Vol 8, pp. 1241-1268, 1988.

[3] Yoganathan, A. P.,Concoran, W. H. and Harrison, E. C., "In Vitro Velocity Measurements in the Vicinity of Aortic Prostheses," *J. Biomechanics.*, Vol 12, pp. 135-152, 1979.

[4] Yoganathan, A. P.,Concoran, W. H. and Harrison, E. C., "Pressure Drops Across Prosthetic Aortic Heart Valves Under Steady and Pulsatile Flow," *J. Biomechanics*, Vol 12, pp. 153-164, 1979.

[5] Figliola, R. S. and Mueller, T. J., "On the Hemolytic and Thrombogenic Potential Occluder Prosthetic Heart Valves from In-Vitro Measurements," *J. Biomech. Engng.*, Vol 103, pp. 83-90, 1981.

[6] Peskin, S. C., McQueen, D. M., "Modeling Prosthetic Heart Valves for Numerical Analysis of Blood Flow in the Heart," *J. Comp. Physics.* , Vol 37, pp. 113-132,1980.

[7] McCracken, M. F., Peskin, S. C. "A Vortex Method for Blood Flow Through Heart Valves," *J. Comp. Physics.* , Vol 35, pp. 183-205, 1980.

[8] Peskin, S. C., McQueen, D. M., "A Three-Dimensional Computational Method for the Blood Flow in the Heart," *J. Comp. Physics.* , Vol 81, pp. 372-405, 1989.

[9] Tarbell, J. M., Gunshinan, J. P., Geselowitz, D.B., " Pulse Ultrasonic Doppler Velocity Measurements Inside a Left Ventricular Assist Device ", *J. Biomech. Engr., Trans. ASME*,, Vol. 108, pp.232-238, 1986.

[10] Mueller, T. J., "Application of Numerical Methods in Physiological Flows," *Numerical Methods in Fluid Dynamics.* , 1978

[11] Idelsohn, S. R., Costa, L. E., and, Ponso, R., "A Comparative Computational Study of Blood Flow Through Prosthetic Heart Valves Using the Finite Element Method," *J. Fluid Dynamics.*, Vol 18, No 2, pp. 97-115, 1985.

[12] Stevenson, D. M., Yoganathan, A. P., and Williams, F. P., " Numerical Simulation of Steady Turbulent Flow Through Trileaflet Aortic Heart Valves-II. Results on Five Models," *J. of Biomechanics,* Vol. 16, Num. 12, pp. 909,926, 1985.

[13] Rosenfeld, M., Kwak, D. and Vinokur, M., "A Fractional Step Solution Method for the Unsteady Incompressible Navier-Stokes Equations in Generalized Coordinate Systems", *J. of Comp. Physics,* Vol. 94, No 1, pp. 102-137, 1991.

[14] Kiris, C., "Computations of Incompressible Viscous Flows Through Artificial Heart Devices With Moving Boundaries," Stanford University, 1991.

[15] Rogers, S. E., Kwak, D. and Kiris, C., "Numerical Solution of the Incompressible Navier-Stokes Equations for Steady-State and Time-Dependent Problems," AIAA *Journal*, Vol 29, no. 4, pp. 603-619, 1991.

[16] Benek, J. A., Buning, P. G. and Steger, J. L., "A 3-D Chimera Grid Embedding Technique," AIAA Paper No. 85-1523, 1985.

[17] Pedley, T. J., and Stephanoff, K. D., "Flow Along a Chanel With a Time-Dependent Indentation In One Wall: The Generation of Vorticity Waves", *J. Fluid Mech.*, Vol. 160, pp. 337-367, 1985.

[18] Ralph, M. E. and Pedley, T. J, "Flow in a channel with a moving indentation ", *J. Fluid Mech.*, Vol. 190, pp. 87-112, 1988.

[19] Baldwin, J. T., and Tarbel, J. M., "Mean Flow Velocity Patterns Within a Ventricular Assist Device ", A.S.A.I. Transactions, Vol. 35, pp. 425-433, Sept. 1989.

N92-22465

COMPUTER INTERFACES FOR THE VISUALLY IMPAIRED

Gerry Higgins
NASA Marshall Space Flight Center
Mission Analysis Division EO41
Huntsville, AL 35812

ABSTRACT

Information access via computer terminals is an essential part of many jobs. This concept extends to blind and low-vision persons employed in many technical and nontechnical disciplines. This paper details information on two aspects of providing computer technology for persons with a vision related handicap. The first is research into the most effective means of integrating existing adaptive technologies into information systems. This will detail research that has been conducted to integrate off-the-shelf products with adaptive equipment for cohesive integrated information processing systems. Details are included that describe the type of functionality required in software to facilitate its incorporation into a speech and/or braille system. The second aspect is research into providing audible and tactile interfaces to graphics based interfaces. The paper includes parameters for the design and development of the Mercator Project. This project will develop a prototype system for audible access to graphics based interfaces. The system is being built within the public domain architecture of X-Windows to demonstrate that it is possible to provide access to text based applications within a graphical environment. this information will be valuable to suppliers of ADP equipment since new legislation requires manufacturers to provide electronic access to the visually impaired.

INTRODUCTION

Over the past several years, computer interfaces have become a major topic of discussion, research and disagreement. The interfaces that exist play a major role in defining how we use computing environments for engineering, analysis, business and numerous other tasks. These interfaces are most commonly discussed and defined in terms of "look-and-feel." This is to say that the appearance of the screen and the response to pointing devices and/or keyboard input determine how users judge the effectiveness of the interface.

The generic terminology, "look-and-feel," emphasizes a visual approach in defining the software user interface. Objects are displayed on the screen to represent textual and graphical information. The relationship of these objects to one another is used to focus the user's attention, define software status, depict options and of course convey meaning. The "feel" of the software is somewhat misleading since there is rarely a tangible tactile element to the user interface. The "feel" relates more to how the visual elements of the display are modified through the use of keyboard or pointing device input.

This emphasis on visual representation has dramatic consequences when a visually impaired individual requires access to computers. The information on the screen must be made available either through an audible output device or through a tactile device. This means a great deal more than just having the screen read aloud by a voice synthesizer. The audible or tactile interface must provide the same explicit and implicit information as is present in the original visual interface. There must be methods for the user to truly interact with the computer just as a sighted counterpart would do. This means audible or tactile responses to inputs, the ability to "look" around the screen, the ability to determine relationships between data that is grouped together and the ability to understand the current state of the software.

In the 1980's, many advances were made in making PC environments available to the visually impaired. However, with one notable exception, these interfaces rely on character based visual environments such as the one found in the early IBM PC. The character based nature of these computing platforms were key to the early success of these endeavors. However, modern technology has rapidly moved computing software and display methodologies into a graphics environment. New technologies must now be developed to insure that visually impaired individuals can continue to participate in tomorrow's information environments.

This paper will outline several of the existing technologies and explain how they can be integrated to make an audible and tactile interface. The latter half of the paper will define research funded through NASA to extend this type of interface into newer graphical environments. This research and prototype is called the Mercator Project and is being developed for the X-Windows Graphical User Interface (GUI) environment.

EXISTING TECHNOLOGIES

Speech and braille output from computers dates at least back to the 1970s. Early speech systems were developed on CP/M machines and HP computers. Soon after the arrival of both the Apple and IBM PC, speech products became available for those environments as well. These speech systems relied on two main components. The first is the speech synthesizer. The synthesizer has the capability of stringing together phonemes to create words. These devices respond to ASCII data that is sent to them from the computer. The second component is now termed a screen access program. This program operates to join the computer with the voice synthesizer in creating audible output from standard off-the-shelf software. This means that through this combination of speech synthesizer and screen access program, a visually impaired user can operate software such as word processors, spreadsheets, data base programs and communications packages.

The speech synthesizer is used to announce each input that the user makes through the keyboard. It also announces information that is displayed on the computer screen. This sounds like a relatively simple solution but the application of the technology proves to have many more elements than may be initially assumed. Some of these elements include:

- How often does the user want to hear an updated screen?

- Is the information being input by the user actually going into the location intended by the user? (Remember, the user cannot just look at the screen to answer this question.)

- As various parts of the screen are updated, what information should be immediately vocalized and what information should be available upon user request? For example, the user needs to know when a menu appears on the screen. However, the user of a word processor will not need to hear the line and column updates after each typed character.

- How does the user know what is in a dialogue box and what is outside the box?

- How does the program differentiate between actual cursor position, highlighted menu options, multiple text windows on the screen and error messages?

- How does the screen access program provide "look-around" capability to the speech user without interfering with the normal operation of the program?

- If colors and video attributes are an important part of the display, how will this information be represented to the speech user?

These issues have been adequately addressed in numerous commercial products that work with character based software running in the MS/DOS environment. Many of the same issues have been addressed to create a speaking interface for the Macintosh operating system. This screen access program, developed by Berkeley Systems, was the first to provide the visually impaired access to a graphical user interface. However, this program does not work with all Macintosh software and only works with text based applications. IBM is working on a version of their Screen Reader program that will provide access to Presentation Manager and OS/2. This is the current IBM graphical user interface.

Tactile devices for producing braille representations of the computer screen are available for MS/DOS computing environments. These devices provide a small window of braille characters, usually one line at a time. This window can be adjusted and moved around the display to show various elements of any display. These devices are limited to functioning only in character based environments.

Both the speech based screen access program and the braille devices rely on the character based nature of the PC platform. In standard DOS applications, a programmer causes a text string to be placed on the video display by either making calls to the Basic Input/Output System or by putting the characters directly into the display memory map. The current adaptive technology can either intercept these calls to BIOS or look directly at the memory map for display information. For each character on the screen, there are 2 bytes in memory that represent the displayed value. One of these bytes is the actual character. The other byte represents information on foreground color, background color and other video attributes such as blinking video. In text mode, it becomes rather easy to obtain information about what is displayed at any given time. Some programs have also become quite adept at analyzing the memory map and informing users about the changes that are important.

This type of technology solved many problems for the visually impaired computer user until software developers began developing more complex and unique forms of conveying interface information. Many developers stopped using cursor positioning as a method to draw the user's attention to a particular part of the screen. This is now routinely done with alternate symbols from the upper part of the ASCII symbol set or by changing colors and video attributes. Screen access programs have become much more sophisticated in providing audible representations by allowing tracking of these newer forms of representations. In research done at NASA on these types of character based interfaces, it was determined that careful planning of a few interface elements of standard software could make access easier for the visually impaired. These elements include:

- Consistent use of colors or video attributes.
  This means that the developer should use a unique combination of colors or attributes to indicate points of interest within a given program. These points of interest might be highlighted menu items, current field names or error conditions.

- Consistent differentiation between elements of the display.
  This implies that there should be variants in color or attributes for different elements of the display. For example, the developer should not use the same video scheme for an error message and a menu selection item.

- Boxes or windows that are formed around text groupings should be complete.
  The current screen access programs are capable of tracing boxes that pop onto the screen but the lines must be complete and drawn with the extended ASCII symbol set.

- Designs that preclude multiple writes of the same information.
  Some DOS based programs and many mainframe programs tend to update the same information multiple times. This is usually because the design of the program doesn't adequately control the order of displayed information. The effect is that the audible output devices speak certain lines multiple times.

- Consistent layout of menus and dialogue boxes.
  Some programs pop up menus in different locations based on information that may not be readily apparent to the casual software user. A menu may appear in the top left corner of the screen while the next menu appears in the middle of the screen. It is much more straightforward for the user of the audible interface if menus consistently appear in the same region of the display.

- Keyboard shortcuts should be available for all actions.
  Visually impaired users of software are still not able to use pointing devices. A programmer should provide keyboard shortcuts for the non-mouse user in order to avoid numerous keystrokes to position an on-screen pointer.

- Consistent usage of cursor positioning.
The cursor is often locked off screen during the entirety of some applications software. In other applications, the cursor's positioning on the screen has little to do with the action or input that is expected from the user. Many screen access program and especially braille displays rely on cursor positioning for basic tracking information. When possible, the cursor or a clearly defined alternate should appear at the display item of highest interest.

Transition

Designers and users of software are becoming more interested in the interface characteristics offered in graphical environments. This means that fewer programs will be introduced with character based displays. When there is no character based display, there is no memory map for the screen access program to use in providing an audible or braille interface. In these graphical environments, text is drawn on the screen by selectively turning on and off display pixels. This allows a lot of flexibility to the sighted user in creating "better looking" displays. However, it presents a tremendous access barrier to non-sighted computer users. This is because there is not an exact representation of the screen image kept in memory. In order to obtain access to the displayed information, software must be developed that intercepts data before it is changed to a graphic image. Other solutions involve performing optical character recognition processes on the display image. This is a costly and slow process at best.

The transition to these types of graphical displays has already begun. It is possible that the computer technology that has been so revolutionary in providing independence to many non-sighted individuals will change to environments that are not suitable for nonvisual usage. The effect of this will be loss of jobs, loss of information access and several steps backward in efforts to integrate the visually impaired into mainstream society.

THE MERCATOR PROJECT

Over the past 2 years it has become especially clear that blind employees at the Marshall Space Flight Center will need to be able to access graphical user interfaces to perform their jobs. Many of the software and hardware systems that are being upgraded for use in space Station support will be based on graphical displays. Analysis shows that the majority of these systems will be hosted in environments that rely on the X-Windows protocol for display of textual and graphics information. The impact of this type of graphical user interface is not limited to work at MSFC. X-Windows is being employed throughout the government, academia, and commercial software implementations.

As a part of research into this problem, MSFC began working with software developers at the Georgia Institute of Technology. The research team headed by Elizabeth Mynatt and including David Burgess, Keith Edwards, John Goldthwaite, Bill Putnam, Tom Rodriguez and Enian Smith has developed a concept and system design for an audible interface to GUIs. This environment is called "The Mercator Project: A Nonvisual Interface for X Windows and UNIX Workstations."[1] The

1. Elizabeth Mynatt and W. Keith Edwards, The Mercator Project, unpublished, 1991.

name refers to "Gerhardus Mercator, a cartographer who devised a way of projecting the spherical Earth's surface onto a flat surface with straight—line bearings."[2] The relationship between Gerhardus Mercator and the Mercator environment is in the mapping of a visual interface to an audible interface. In addition, all interfaces can be said to aid in navigation through the underlying software. The Mercator environment will be prototyped during the winter and spring of 91-92 at the Georgia Institute of Technology. NASA personnel including Gerry Higgins and Craig Moore will be providing technical expertise, requirements and design concepts. The initial prototype will be hosted on a Sun Workstation. However, the intent from the outset of the project is to make the Mercator environment platform independent. The goal is to make a final system that will work in any standard UNIX implementation on a workstation that supports X-Windows.

The MOTIF interface standards are being employed to help insure that Mercator will work with a wide variety of applications. MOTIF is an interface standard that is expected to be used by a majority of software developers who program X-Windows applications. In the Computer Glossary, Motif is defined as: "(Open Software Foundation/Motif) A graphical user interface (GUI), developed by OSF, that offers a PC-style behavior and appearance for applications running on any system that supports X Window, Version 11. It conforms to POSIX, ANSI C and X/Open's XPG3 standards."[3] By concentrating on MOTIF applications, it is possible to define specific interaction characteristics that can be expected within the Mercator/MOTIF environment. These include:

- Predefined keyboard shortcuts.
- Standard interface objects.
- A limited number of menu types, 3 to 4 basic menus.
- Predefined method of displaying visual queues.
- Eighteen different types of cursors.

The predefined keyboard shortcuts within MOTIF will facilitate easier use of the environment to the nonvisual user because the user will not need to rely on a pointing device. X-Windows applications generally rely heavily on navigation through movement of a mouse. This type of mouse movement works well for the visual user but is especially tricky for the visually impaired user because there is no fixed point of reference. This means that there is no relationship between the physical position of a mouse on the desktop and the pointing cursor that is on the screen. There are two approaches to solving this problem within Mercator. The first is to rely on the keyboard shortcuts as a way of activating menu items and/or scroll bars. The second method is to employee a pointing device with a fixed point of reference. For this approach, the research team is investigating the use of a touchpad. The touchpad provides a physical relationship between a position on a tactile surface and the display

2. Elizabeth Mynatt and W. Keith Edwards, The Mercator Project, unpublished, 1991.

3. Alan Freedman, The Computer Glossary (The Electronic Version), 1991.

objects on the screen. It should be possible to lay a raised grid on top of a commercially available touchpad in order to give the visually impaired user basic positional information.

In X-Windows, display is accomplished through a set of widgets that handle items such as buttons, scroll bars, text fields, and other icons. MOTIF offers a standard set of widgets to be used in applications. These widgets are used as display objects that must be interpreted into the Mercator environment. Mercator will be designed to provide audible access to the standard display objects available through MOTIF. Mechanisms will be available to provide audible differentiation between these widgets. There will be descriptive methods to let the Mercator user know the type of display widget that is being utilized by the application. This will be done through a combination of voice queues and representational sounds. This ability to indicate the type of display widget will extend into areas such as display cursors and visual queues. The Mercator user will need the same ability as the sighted user to manipulate these display objects and to know the effect that the manipulation is producing. The Mercator environment will provide this type of feedback through speech output, audible tones, and responses to input from the touchpad.

Another feature that is being developed to help in orientation to manipulation of these display objects is a 360 degree sound environment. To use this feature, the user will ware a set of stereo headphones. Sound objects will be audibly arrayed in a 360 degree spectrum around the user's head. These objects will move within the sound space as the user manipulates them via the keyboard or touchpad. Additionally, it will be possible to differentiate between sound objects through pitch and tone of voices or audible queues.

Development Approach

A two tiered approach will be taken in development of the Mercator environment. The first tier is the applications level. This development will deal with making standard X-Windows and MOTIF applications available in the Mercator environment. The development will concentrate on only text applications. The second tier of the development is to create an audible workspace that is analogous to the visual desktop seen in many graphical interfaces. This element of the Mercator environment will not be required if the user is only going to run standard X-Windows applications.

Basic X-Windows access

X-Windows is a graphical environment that offers a lot of possibilities in creating a nonvisual interface. This is because of the client/server nature of the X-Windows protocol. The client, a piece of software running on either the workstation or a networked computer, must communicate with a server about the information to be displayed and the input provided by the user. This is called interprocess communications.[4] This requires messages to be sent between the client and the server. This message passing is one of the primary means that will be employed to obtain information for the Mercator audible display. As data is past back and forth between the client and server, the

4. Robert W. Scheifler and J. Gettys, The x window system, ACM Transactions on Graphics, (2), April 1986.

Mercator software will track displays as they are being built and manipulated.[5] This type of additional processing will not interfere with the message passing between the client and server that was intended by the software developer. It will only act to intercede on behalf of the user of the nonvisual interface. This "tap" into the flow of information between client and server will allow a partial off-screen model of the display to be built. This off-screen model will be used by the Mercator environment to produce the audible element of the nonvisual display. This model will also be used to track input by the user such that the Mercator environment will know how to provide audible queues as to the changes on the display and actions that might be taken. This client/server relationship is very beneficial because it restricts developers from directly manipulating display elements on the screen. All displays must occur through the pipeline between the client and X-Windows server. This circumstance makes it much more likely that the Mercator can provide full access to a wide variety of software.

The latest release of X-Windows, X11R5, provides new functionality that will be very useful in building this access tool. This release allows clients to be queried about the state of display objects that relate to a given application. Through this query process, Mercator will be better able to determine how display images change, how pointers and cursors have been altered, and the actions that are available from pull down menus or other hidden elements of the display.

Mercator extensions

The desktop metaphor is used to provide sighted users of graphical interfaces with a familiar working environment. These environments provide methods for displaying objects, moving objects around, transferring information between objects and launching applications. These types of activities are also performed by visually impaired individuals but the activities are usually accompanied by tactile information. For example, picking up a physical document (reinforced by the touch of the document) and then locating the physical trash can for document deposit. Since there is no immediate economical way of reproducing this tactile information, the Mercator environment will provide feedback to these types of activities via voice output. It will be necessary to vocalize information about what the objects are and the types of actions that can be associated with the objects.

The Mercator environment will provide an extension to the desktop metaphor based on research done at Xerox PARC.[6] The concept derived by this research is termed Rooms. In the Rooms metaphor, the user can collect like objects or applications to perform similar tasks. This helps to avoid the clutter that can be found on the proverbial desktop. These Rooms or workspaces can be connected via doors to form a network of workspaces that can be arranged to suit the users needs. This is an especially appropriate metaphor for visually impaired users. Most non-sighted computer users are very capable of visualizing themselves in the midst of this workspace. It is appropriate to think of oneself in the middle of an environment with the tools needed to do a particular job arrayed around a room. Mercator will provide appropriate audible feedback to represent moving around this workspace and from one workspace to another as in audible doorways. It should be possible for the user to learn to navigate through this network in much the same way that mobility skills are taught

5. Elizabeth Mynatt and W. Keith Edwards, The Mercator Project, unpublished, 1991.

6. D. Austin Henderson and Stuart K. Card, Rooms: The use of multiple virtual workspaces to reduce space contention in a window-based graphical user interface, ACM Transactions on Graphics, pages 211-243, July 1986.

to non-sighted individuals. The 360 degree sound capability, mentioned earlier, will be a key ingredient to producing these Room effects. This technology will allow for positional representations that can be acted on by the listener. Navigation effects that produce pathway information can be simulated for passing from one room to another within the 360 degree sound space.

Prototype Completion

The Mercator prototype is scheduled to be completed by March of 1992. It is hoped that additional funding can be found to produce the full Mercator environment. This of course depends on results obtained from the initial work. Georgia Tech has been assembling a candidate list of visually impaired computer users to view the system after prototype completion. Their evaluation of the work will influence future development of the system. Additionally, many individuals who participate in providing adaptive solutions are being asked to make input to the design and development of the system. It is the goal of this project to ensure that visually impaired individuals continue to have access to work place information and that the users of Mercator be able to continue to work along side their sighted counterparts.

N92-22466

EXTENDED ATTENTION SPAN TRAINING SYSTEM

Alan T. Pope
Human Engineering Methods Group
Human/Automation Integration Branch
NASA Langley Research Center
Hampton, VA 23665

Edward H. Bogart
Lockheed Engineering and Sciences Company
Hampton, VA 23665

ABSTRACT

Attention-deficit Disorder (ADD) is a behavioral disorder characterized by the inability to sustain attention long enough to perform activities such as schoolwork or organized play. Treatments for this disorder include medication and brainwave biofeedback training. Brainwave biofeedback training systems feed back information to the trainee showing him how well he is producing the brainwave pattern that indicates attention. The Extended Attention Span Training (EAST) system takes the concept a step further by making a video game more difficult as the player's brainwaves indicate that attention is waning. The trainee can succeed at the game only by maintaining an adequate level of attention. The EAST system is a modification of a biocybernetic system that is currently being used to assess the extent to which automated flight management systems maintain pilot engagement. This biocybernetic system is a product of a program aimed at developing methods to evaluate automated flight deck designs for compatibility with human capabilities. The EAST technology can make a contribution in the fields of medical neuropsychology and neurology, where the emphasis is on cautious, conservative treatment of youngsters with attention disorders.

ATTENTION MANAGEMENT AND AUTOMATION

The management of attention is important for success in a wide range of endeavors from sport to study. The ability to remain aware of fluctuating attentional states, the ability to maintain effective states, and the ability to recover efficiently from attention lapses are valuable in task settings requiring recognition and response. In addition to personal differences in these abilities, there are also differences in the nature of tasks that either foster or discourage effective attention. The advertising industry has developed a technology of manipulating attention and awareness using presentation characteristics such as pacing, sequencing, complexity and color.

As automated systems become more capable, human operators spend less time actively controlling such systems and more time passively monitoring system functioning. This type of task demand challenges human capability for sustained attention.

Attention may be usefully characterized by three aspects: 1) distribution (diffused versus concentrated); 2) intensiveness (alert versus inattentive); and 3) selectivity (the "what" of attention) [3]. Distribution and intensiveness are influenced by the state of awareness being experienced, and selectivity refers to the contents of awareness. In designing for effective integration of human and system, it is important to provide ready access to useful information so that the contents of awareness support informed action. It is also important to design for human involvement in system function to promote effective states of awareness; i. e., to promote consistent mental engagement in the supervisory task. Mental engagement in automated environments may be enhanced by judicious allocation of task responsibility between the human and the automated system.

The Extended Attention Span Training System (EAST) system is a modification of a biocybernetic system that is currently being used to assess the extent to which automated flight management systems maintain pilot engagement.

A description of the biocybernetic system will provide the background for understanding how an adaptive task regulated by electrical brain activity forms an environment for training attention skills.

A BIOCYBERNETIC ASSESSMENT SYSTEM

The Automated Task Environment

A methodology and system have been developed to support the determination of optimal human (manual)/system (automated) task allocation "mixes," based upon a brain activity criterion of consistent mental engagement. In this methodology, an experimental subject interacts with a set of tasks presented on a desktop computer display while the subject's electrical brain activity is monitored.

This evaluation methodology has been developed in the context of computer-assisted flight management. The tasks in the set, the Multi-Attribute Task (MAT) Battery [2], are designed to be analogous to tasks that crew members perform in flight management. The MAT Battery display, depicted in Figure 1, is composed of four separate task areas, or windows, comprising the monitoring, tracking, communication, and resource management tasks. Each task may be fully or partially automated. The monitoring subtask (upper left window) requires a subject to detect warning light changes and scale pointer offsets. When this subtask is automated, these events are responded to by the computer. The compensatory tracking task (upper middle window) requires a subject to keep a moving symbol within a central rectangle. This task is automated by constraining movement of the symbol in one or both axes to within the central rectangle. The communications task (lower left window) requires a subject to discriminate an auditory message intended for his flight from other messages and to follow the message command to set radio frequencies. When this subtask is automated, the frequencies are set automatically when the message is received. The resource management subtask (lower middle window) is presented to subjects as a fuel management task. A subject is required to maintain a prescribed level in each of two tanks by controlling pumps from reservoirs. The computer maintains the levels when the task is automated.

The level of automation of the task battery may then be varied so that all, none, or a subset of the system control functions require subject intervention. This variable automation feature enables a range of levels of demand for operator involvement in system management to be imposed. Corresponding to these levels of involvement are degrees of mental engagement that are often spoken of in terms of being "in" or "out of the loop." Mental engagement in an automated task may not be sufficient to promote an effective state of awareness. Monitoring brain activity provides a window through which to view the experiential states of a subject in this situation.

The Electroencephalographic Engagement Index

The electroencephalogram (EEG) or brainwave has long been used to index states of consciousness or awareness. Stages of sleep are readily mapped by analyzing the frequency content of the electrical activity of the brain. Less well studied are the stages of waking consciousness. However, recent work [8] has identified characteristic patterns in three established brainwave frequency bands that distinguish among various states of attention. Relatively greater beta (13-20 Hz) activity has been observed for vigilant states, whereas alpha (8-13 Hz) activity predominates in alert but less mentally busy states, and theta (4-8 Hz) activity rises as attention lapses. These brainwave-state correspondences have proven useful in assessing attention-related disorders [5] as described below. However, within these guidelines there are significant individual differences. The set of frequency bands and recording sites that discriminate best among states of awareness for one individual are not likely to be the same set for another. Therefore, to derive an EEG index of attention with which to assess mental engagement in a task, it is necessary to develop each subject's EEG-to-state mapping profile.

This profiling procedure is conducted in the task environment described above. The procedure involves recording topographical EEG data from 19 electrode sites while a subject performs the MAT battery. The MAT is systematically stepped through its range of automation levels. At each step, a spatial map of EEG activity is generated for each of the alpha, beta, and theta frequency bands that portrays the scalp distribution of activity within that band. In this way, a trio of brainmaps is generated that corresponds to the state experienced at each level of task automation. The EEG data represented in the entire set of brainmaps are used to derive state discriminant functions for the subject.

Once an individual subject's characteristic mapping profile has been determined, an index is constructed which is designed to be maximally sensitive to changes in state. The index has the form:

$$(K_1 * bS_1)/(K_2 * aS_2 + K_3 * tS_3)$$

where bS_1 = beta activity at electrode site 1
aS_2 = alpha activity at electrode site 2
tS_3 = theta activity at electrode site 3

with S_1, S_2, and S_3 representing the sites at which the activity in the associated band is most discriminating among states and K_1, K_2 and K_3 are coefficient weights that reflect the relative contributions that the three band/site factors made to the discrimination. This index is constructed to have higher values for subject states corresponding to greater degrees of mental engagement; i. e., greater demands for operator involvement.

The Adaptive Task Concept

The engagement index is next employed with the MAT battery in a closed-loop control paradigm to observe the effects of adaptively allocating task responsibility between operator and automated system (Figure 2). That is, the task battery is adapted to the subject's degree of engagement in the task by assigning an additional subtask to the subject when it is determined that task engagement is waning over a time interval (Figure 3). Conversely, when the engagement index exhibits a sustained rise, indicating that the subject is capable of monitoring attentively, an additional battery subtask is automated. In this way, the feedback system eventually achieves a steady-state condition in which neither sustained rises nor sustained declines in the engagement index are observed. The combination of automated and manual subtasks, the task "mix," that the subject is presented in this condition may be considered optimal by the criterion of mental engagement reflected in the EEG state index.

This adaptive process is essentially a feedback control process whereby the reference EEG condition, stable short-cycle oscillation of the engagement index, is achieved by systematic adjustment of task demand for operator participation. EEG parameters have been used similarly as control variables in the critical administration of anaesthesia. In that application, closed-loop feedback control methods compare the set-point of the control variable, e. g., median EEG frequency, with the value actually measured to modify rate of drug delivery [7].

The adaptive system is designed to evaluate automated task environments to determine the requirements for operator involvement that promote effective operator awareness states. However, the assessment procedure may function as a training protocol in that the subject is rewarded for producing the EEG pattern that reflects increasing attention level by having the automated system share more of the work. With practice, a subject may learn how to deliberately control subtask allocation to the level at which he prefers to work. This observation led to consideration of the adaptive task concept for an attention training application.

THE EXTENDED ATTENTION SPAN TRAINING (EAST) SYSTEM

Biofeedback Training For Attention-deficit Disorder

Symptoms of pure Attention-deficit Disorder (ADD)--short attention span and poor focusing and concentration skills--are present in a majority of children who are hyperkinetic, or who have a specific learning disorder or who exhibit a conduct disorder [1] [5]. Lubar [5] and coworkers have found that topographic brain mapping of EEG frequency bands discriminates between children classified as pure ADD and matched controls, and that the discriminations are stronger during task performance than during baseline conditions. ADD children showed greater increases in theta activity and decreases in beta activity during tasks. These findings strengthened the rationale for providing EEG biofeedback training to reverse these brain activity trends and improve attentional abilities.

Training is accomplished by providing the student with a real-time display of the levels of beta and theta activity being produced. The display serves as both information and reward for the child's efforts to reduce theta activity and increase beta activity. Success at producing the desired brain activity changes as well as improvements in psychometric performance have been reported as outcomes of training [5].

Mulholland [6] described a system for visual attention training using alpha activity as the control variable. Lubar [5] specifically recommends the use of theta/beta and alpha/beta ratios similar to the engagement index described above as sensitive discriminators of ADD. The Extended Attention Span Training (EAST) system uses brain activity band ratios in a training protocol that also employs the adaptive task concept.

The Attention Training Video Game

The MAT environment can serve as an attention training system as explained previously. However, the mission of achieving uneventful flight is likely to be less motivating to youngsters than winning at a dynamic, competitive video game. The EAST system is a prototype game environment intended to demonstrate the application of the adaptive task concept in a training format appealing to children. The video game platform selected presents a space battle scenario in which the player pilots a fighter ship in an attempt to reach and explode an enemy base while warding off attacks from the base's defenders. The action is viewed through a sighting window with cross-hairs. Other data such as range to targets, speed and weapons status are presented onscreen (Figure 4). A joy-stick is used to acquire and fire upon targets.

For the EAST application, the game is reprogrammed to impose an additional criterion for success. The game is made virtually impossible to win due to the evasive maneuvering and overwhelming number of the defenders. However, as a player concentrates and focuses better, the defenders maneuver less and there are fewer of them. Therefore, as the player does better at maintaining attention, the manual part of the game is made more manageable. The player learns that winning is contingent upon maintaining attention even as the game becomes less compelling. The attentive state is explained to the player as a special mental power, that is, being attentive is invoking this power to aid in winning. This is accomplished by programming the number and movement of targets to be a function of an attention-sensitive EEG band ratio. The EEG signal is sensed at a single scalp site and conditioned by an inexpensive isolated interface [4] into the game computer serial port. Frequency analysis, ratio calculation and game control are accomplished in software.

The next phase of work will evaluate the benefits of the adaptive task concept for attention training with a clinical population. It may be that the most effective use of the EAST system would be to augment current biofeedback paradigms for Attention-deficit Disorder by providing a rewarding environment in which trainees can demonstrate and improve skills learned in prior training. To offset the effects of ADD, EEG biofeedback must be integrated into an academic skills program including reading, mathematical skills and attention skills. The EAST technology represents a prototype of a new generation of computer game environments that teach valuable mental skills beyond eye-hand coordination. The technology can make a contribution in the fields of medical neuropsychology and neurology, where the emphasis is on cautious, conservative treatment of youngsters with attention disorders.

REFERENCES

1. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, Third Edition - Revised. Washington, D.C.: APA, 1987.
2. Arnegard, R. J. and Comstock, J. R., Jr. Multi-Attribute Task Battery: Applications in Pilot Workload and Strategic Behavior Research. Proceedings of the Sixth International Symposium on Aviation Psychology, Columbus, Ohio, 1991, pp. 1118-1123.
3. Carver, C. S. and Scheier, M. F. Attention and Self-Regulation: A Control Theory Approach to Human Behavior. Springer-Verlag: New York, 1981, pp. 34,35.

4. Ciarcia, S. Computers on the Brain. Byte, 13, no. 6: pp. 273-285, 1988.
5. Lubar, J. F. Discourse on the Development of EEG Diagnostics and Biofeedback for Attention-Deficit/Hyperactivity Disorders. Biofeedback and Self-Regulation, 16, no. 3: 201-225, 1991.
6. Mulholland, T. B. Training Visual Attention. Academic Therapy, 10, no. 1: 5-17, 1974.
7. Schwilden, H., Stoeckel, H., and Schuttler, J. Closed-Loop Feedback Control of Propofol Anaesthesia by Quantitative EEG Analysis in Humans. British Journal of Anesthesiology, 62: 290-296, 1989.
8. Streitberg, B., Rohmel, J., Herrmann, W. M., Kubicki, S. COMSTAT Rule for Vigilance Classification Based on Spontaneous EEG Activity. Neuropsychobiology, 17: 105-117, 1987.

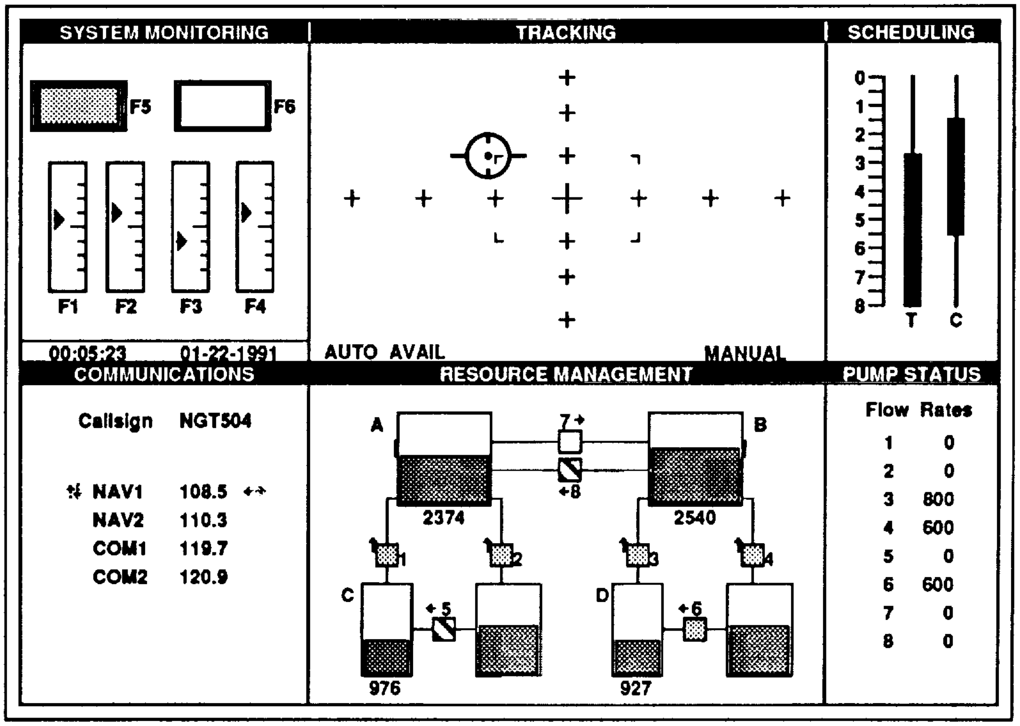


Figure 1. Multi-Attribute Task Battery Display

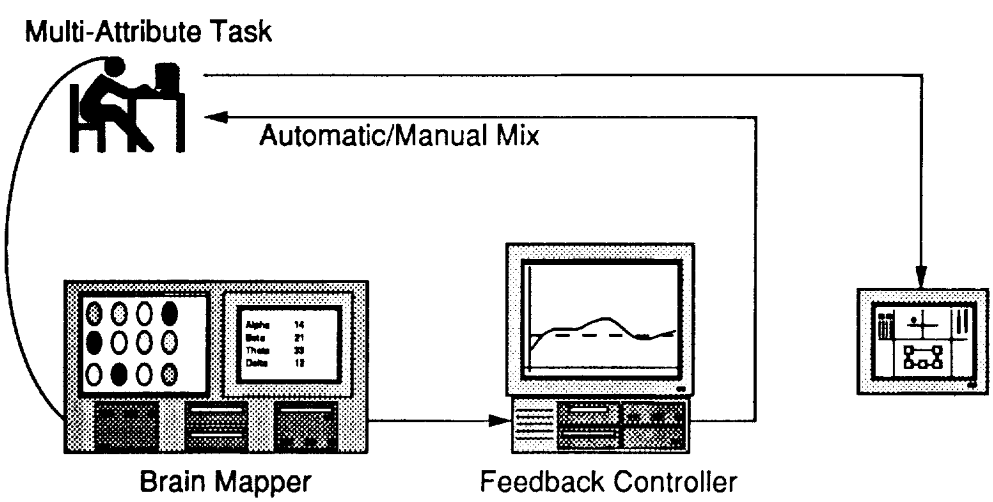


Figure 2. State-Contingent Adaptive Task Environment

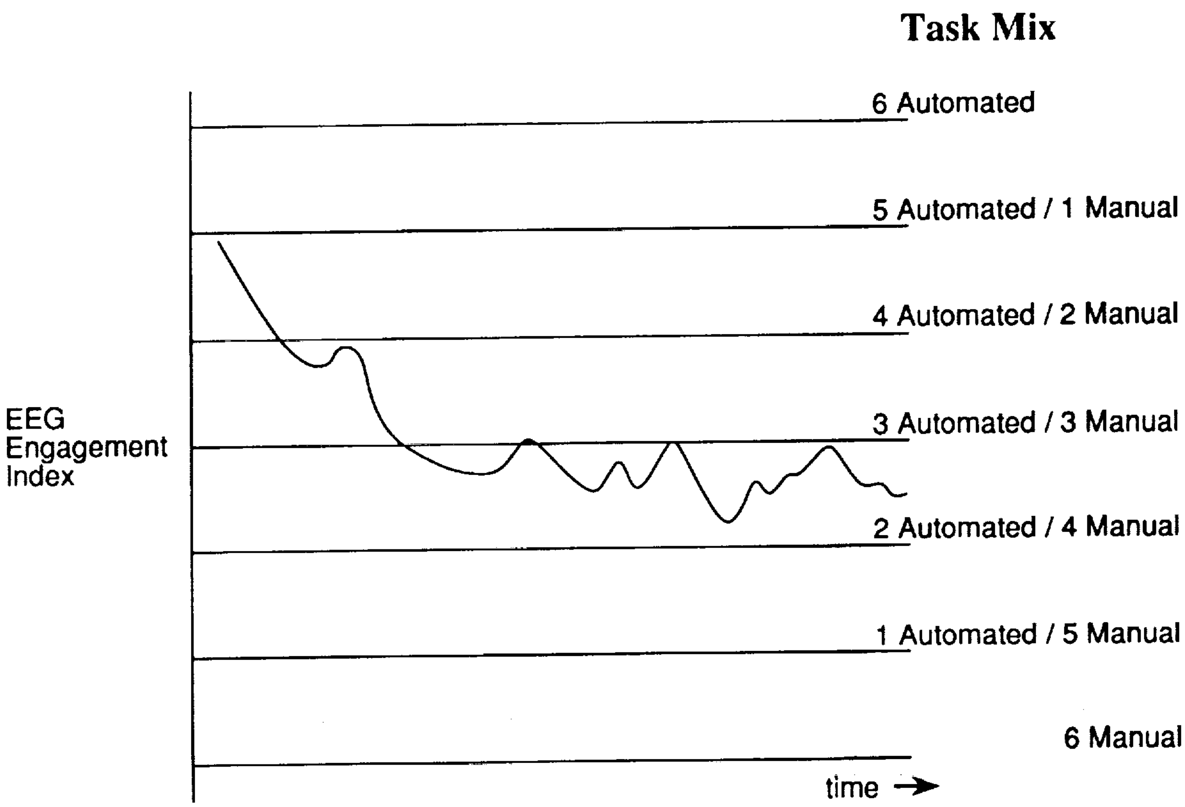


Figure 3. State-Contingent Adaptive Task Behavior for Automation Assessment

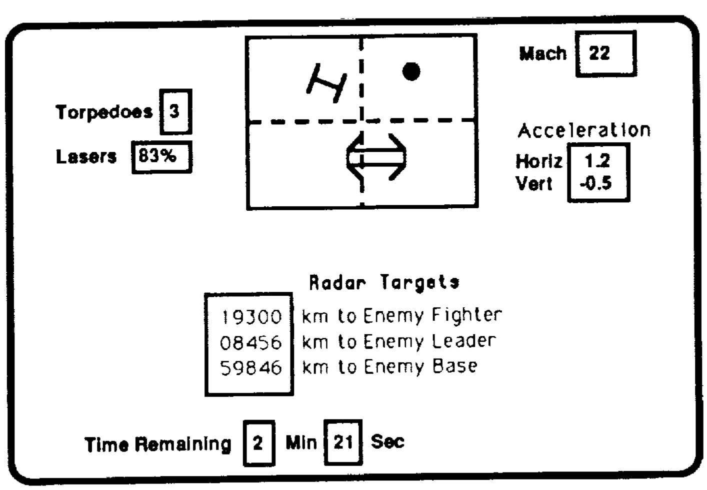


Figure 4. EAST Game Display

N92-22467

MAN/MACHINE INTERACTION DYNAMICS AND PERFORMANCE (MMIDAP) CAPABILITY

Harold P. Frisch
Head/Robotics Applied Research
NASA/Goddard Space Flight Center (GSFC)

ABSTRACT

The creation of an ability to study interaction dynamics between a machine and its human operator can be approached from a myriad of directions. The Man/Machine Interaction Dynamics and Performance (MMIDAP) project, lead by the Goddard Space Flight Center (GSFC), seeks to create an ability to study the consequences of machine design alternatives relative to the performance of both the machine and operator. The class of machines to which this study is directed includes those that require the intelligent physical exertions of a human operator. While GSFC's Flight Telerobotic's program was expected to be a major user, basic engineering design and biomedical applications reach far beyond telerobotics. This paper outlines ongoing efforts of the GSFC and its University and small business collaborators to integrate both human performance and musculoskeletal databases with analysis capabilities necessary to enable the study of dynamic actions, reactions, and performance of coupled machine/operator systems.

INTRODUCTION

Goethe's insightful statement "The pieces I am holding in my hand, what I lack is the clarifying bond" provides an accurate definition of where GSFC's MMIDAP capability is and where it is headed. Since the project's last summary paper [1], "the pieces" have been identified, gaps recognized, and the process of creating "clarifying bonds" initiated. This paper updates the collaborative efforts of participating groups to develop a collection of marketable MMIDAP software products that will be open ended analysis capabilities. These will enable biomechanical and human performance analysis methods to be interfaced with appropriate databases. The open ended structure of the software systems will permit MMIDAP computational tools to keep pace with the state of the art by allowing the incorporation of new analysis and performance evaluation techniques.

The development of a MMIDAP capability requires both anatomical and human performance data. It requires physical testing in both the laboratory and workplace environments. Finally it requires analysis capabilities that will provide insight, understanding and the ability to extrapolate limited laboratory test data to a variety operator population groups and operational environments.

The use of MMIDAP software systems can be viewed from several different perspectives; for example:

o Mechanical and concurrent engineering will see MMIDAP providing design engineers with an ability to take human operator strengths and weaknesses into account in the early stages of machine design. The ability to obtain reliable estimates for machine operator comfort, performance, fatigue, and the potential of machine induced medical trauma during the very early pre-prototype stages of machine design will benefit both manufacturer and operator.

o Biomechanical engineering, sports medicine, orthopedics, and physical therapists, will see MMIDAP providing detail musculoskeletal system response predictions for activities associated with work and recreation.

o Rehabilitation technology service providers will see MMIDAP providing an ability to design mechanical assists and prosthesis devices to improve the quality of life for the handicapped. Furthermore, it will allow machine designers to more effectively account for the needs of handicapped operators early in the design cycle.

o Industrial ergonomists will see MMIDAP providing an ability to better understand the relationship between repetitive work and the onset of discomfort, pain, fatigue, and more serious medical problems. It will also allow job placement specialists to obtain quantifiable measures that can be used to determine if a particular employee (male, female, handicapped) has the physical resources necessary to safely operate a particular machine in the work environment for which it was designed.

JUSTIFICATION

The final report of the 1985 Integrated Ergonomic Modeling Workshop [2] contains a detailed review of pre-1988 software capability along with a list of recommendations for future research. It specifically remarks that "there is a paucity of dynamic interface models" and that "an integrated ergonomic model is needed, feasible, and useful. " The report's review of existing capability demonstrates that ergonomic modeling software has been primarily developed to support aerospace cockpit design, design for product maintainability, and whole body dynamics associated with automotive vehicle crash and pilot ejection. Some work exists under the general heading of optimization of sports motion. However, there is virtually nothing to support designers who must evaluate man/machine interaction dynamics and performance with or without survival gear, in hostile environments on earth, or in the reduced gravity environment of space.

The evolution and integration of engineering and medical sciences are bringing dynamic new problem solving capabilities within reach. A natural step commensurate with advances in analytic power involves the integration of existing knowledge and data. While gaps exist, attempts to integrate the significant pool of knowledge and data currently available would not only yield a useful end product, but would also serve to more precisely identify the location and nature of needs. While the scope of the MMIDAP project emphasizes biomechanical aspects of the human musculoskeletal system, issues are raised regarding interfaces to models of neurologic and cardiovascular systems to facilitate evolution of the basic tools envisioned.

MMIDAP DEVELOPMENT PLANS

The MMIDAP project is being planned at three levels of detail:

1. Integrated biomechanical system response and system performance. A four university team of multi-disciplinary researchers from The University of Texas at Arlington, The University of Iowa, Case Western Reserve University, and The University of Colorado at Denver have recognized the far reaching potential of this capability. A recent report [3] defines a plan for integrating existing anthropometric, musculoskeletal, and human performance databases with existing analysis capabilities. The report considers issues that impact the plan for a computer-based tool of sufficient scope and flexibility to assist both researcher and clinicians; i. e. , physical therapists, rehabilitation technology service providers, etc. Strategies with generic utility and rational for integrating anatomy (structure) as well as function and performance are presented. A standard approach to interfacing models across different levels is considered. Current data availability, gaps in data availability, data warping methods, and acquisition feasibility are reviewed in the context of supporting modeling and analysis functions in areas of practical interest and concern. It is argued therein that the need for and the relevance of this capability is based upon the following observation that: o Significant advances in analytic power can be realized through the integration of existing knowledge and discrete capabilities. o While gaps exist, a significant pool of such knowledge and data is presently available. o Technology is readily available to make implementation and dissemination feasible. o Clinical problems and research questions exist for which it is difficult to foresee solutions without the envisioned capability, thus rendering merit to the objectives. o Integration is unlikely to occur without a coordinated planning process.

2. The development of a software system to support the detailed evaluation of machine operator's task performance from the perspective of whole body response and its correlation with required levels of human performance resources needed for task completion is provided in reference [4].

3. The development of a software system to support the detailed evaluation of machine operator performance from the perspective of musculoskeletal system response and its correlation with the potential for work related pain,

discomfort, fatigue, and the more serious medical problems is provided in reference [5].

These planning efforts have highlighted the need for establishing bonds between collaborators. These bonds are effectively protocalls that allow data collected by one group to be used by another.

VISIBLE HUMAN PROJECT

The National Library of Medicine (NLM) is currently supporting a first project aimed at building a digital image library of volumetric data representing a complete normal adult human male and female [6]. This "Visible Human Project" will include digital images derived from photographic images from cryosectioning, computerized axial tomography (CAT), and magnetic resonance imaging (MRI). Members of the four-university MMIDAP project team are taking a critical view of the state of biomechanical and human performance analysis possibilities vs the state of associated data availability. It will develop a hierarchial tree of biomechanical and human performance analysis capability vs data availability, along with a plan for its realization. The emphasis is to develop the infrastructure for a systematic, engineering approach to solving human system problems and recognizing current limitations.

In reference [3], an initial plan for integrating biomechanical analysis and modeling capabilities, as well as data, into a workstation tool is defined. One problem with designing biomechanical analysis capability for the market place is that of input data development. Even the most sophisticated users find it difficult to define all musculoskeletal and human performance data needed to characterize specific human functions. In reference [6] it is recognized that the "visible human project" should eventually create an ability to both view anatomical cross sections and to query the database using functional queries. The ability to use MRI and CAT data to warp the database would provide an ability to obtain individualized detail anatomical data per subject.

ANTHROPOMETRIC AND MUSCULOSKELETAL DATABASES

A review of currently supported anthropometric data bases and computer models used in the field of ergonomics may be found in reference [2]. A source book for multiple muscle systems and movement organization, along with a survey and listing of human musculotendon actuator parameters from over 20 different published sources [7] is provided in reference [8]. Anthropometric and musculoskeletal data used to support musculo load sharing research carried on at the University of Wisconsin at Madison is provided in reference [9].

One major problem with existing biomechanical data is that they come from so many different sources, with almost as many different measurement reference frames. The NLM's Visible Human Project is presenting the biomechanics community with a unique opportunity to fill data gaps and to obtain a consistent reference source of fundamental biomechanical data.

BASIC ELEMENTS OF HUMAN PERFORMANCE

Biomechanical data alone are not sufficient for man/machine interaction dynamics and performance analysis, human function and human performance is also needed. A review of ongoing work in the quantitative measurement and assessment of human performance is provided in reference [10].

One approach to human performance task analysis is to decompose a task into fundamental components that have an associated set of quantifiable measures of performance. Kondraske refers to these fundamental components as "Basic Elements of Performance" (BEP's). BEP's have measurable attributes that can be databased according to individual and population cross section. BEPs form the basis of an elemental resource model for characterizing all aspects of the human system and explaining their relationship to tasks. As a result of multi-system measurement research, Kondraske discovered the need to first delineate general system performance theoretical constructs. A particularly attractive feature of the elemental resource model is that it is derived from a straightforward set of "first principles" that are generalizable to all systems; e.g., musculoskeletal, central processing, and life sustaining. Each BEP can be viewed as a specific functional unit and one of its dimensions of performance. For example, the functional unit (e.g. knee extensor) and one of its dimensions of performance (e.g. speed) defines a particular BEP.

Dimensions of performance identify unique qualities of performance such as speed, range of motion, accuracy, etc. Together they define a multi-dimensional performance space.

If one views BEP's as resources, then one can introduce the concept of resource availability. For example, once task decomposition into BEP's is complete, human resource requirements can be estimated. Established norms for population cross sections of interest can be used to determine if tasks are within the available resource capabilities of a particular person or of select population groups. Conversely, BEP requirements above the norm can be used to pinpoint exactly where the man machine interface needs improvement. A critical ramification of the resource construct is that it forces consideration of only those dimensions that represent desirable qualities and by definition provides a consistent method for quantifying performance. Thus,confusion resulting from dual concepts that has pervaded the field (strength vs. weakness, endurance vs. fatigue, etc.) is eliminated and a clear modeling framework emerges.

HUMAN PERFORMANCE DATABASE

Reference [11] provides a good overview of task decomposition via BEP's and methods used to database the BEP records of over 3000 patients at the University of Texas at Arlington's Human Performance Institute (HPI). This work originally was focused on the field of Physical Therapy and Rehabilitation Engineering. It is now recognized that HPI's databased information and measurement systems are identical to what is needed to support the MMIDAP project.

Task decomposition into BEP's is an attractive approach to providing a systematic basis for predicting human performance and for extrapolating experimental test data to the machine's operational environment. Unfortunately, the problem of task decomposition is non-trivial. One approach is to make use of the digital image library of the NLM's Visible Human Project. Conceptually it appears possible to create an expert system support shell that would allow users to define a high level task. Output would be a visual display of all anatomical components involved in the task and associated basic elements of performance. It seems reasonable to expect that this capability could be developed as a hierarchial structure. BEP's combine to form simple tasks, simple tasks combine to form more complex tasks, etc. Issues associated with body posture and kinematic redundancies within the human body are not being under estimated. The whole body digital mapping project intends to pursue this concept and to explore feasibility by focusing on the human knee complex in 1991.

MUSCULOSKELETAL SYSTEM MODELING

Central to the effort are the methodologies to be used for biomechanical analysis. The human musculoskeletal system is generally modeled as a mechanical system of links, joints, and actuators. Depending upon analysis fidelity, links are modeled as individual bones that are modeled as either rigid bodies, flexible bodies, or body clusters. Joints can be modeled as simple mechanical hinges or as complex anatomical joints that account for rolling and sliding contact between the irregular contact surfaces. Actuators can be modeled as resultant torque producing motors at each joint or complex force producing elements that attempt to model the musculotendon lines of action between points of origin and insertion. In theory, it is also possible to include such effects as muscle contraction dynamics and muscle recruitment to attempt to predict resultant biosystem dynamics.

General purpose software modeling capabilities exist to support this type of mechanical system modeling. The state of the art of general purpose multibody dynamics modeling is quite mature. However, biomechanics application will demand the development of enhanced capability. Reference [12] provides a reasonably complete overview of international capabilities in multibody dynamics.

Given the availability of these analysis tools, their application requires system characterization data. Kondraske of the University of Texas at Arlington has this nations most extensive database of human performance data. Seirig at the Universities of Wisconsin and Florida has most probably this nations most extensive database of skeletal and musculotendon data compatible with mechanical system analysis. Whitlock and Spitzer of the University of Colorado are under contract to the National Library of Medicine (NLM) to create whole body digital maps of both a human male and female. The MMIDAP project intends to integrate these databases. It is recognized that anthropometric and other biomedical and human performance databases exist, which will be integrated as need.

INVERSE DYNAMICS

Multibody simulation models have been successfully used to model certain classes of musculoskeletal systems. However, modeling weaknesses exist and these must be recognized before one attempts to use multibody tools for general biomechanical application. Dynamic analysis of mechanical systems is dominated by the need to solve the forward dynamics problem; i.e., given a prescribed set of internal and external loads, predict system response. Attempts to perform forward dynamic analysis with musculoskeletal systems is usually stopped by the analyst's inability to mathematically characterize the human's cognitive processes which generate the neural activation signals that stimulate the body's musculo actuator system.

The MMIDAP project recognizes this fundamental limitation. Instead, it concentrates on the inverse dynamics problem. Graphical animation and laboratory testing techniques exist for obtaining an estimate of human dynamic response for a broad range of activities. If sufficient information can be obtained (displacement, rate, acceleration, and external loads), then inverse dynamics methods can be used to predict what the resultant musculo actuator loads had to be to produce the defined input data. These results can then be compared with known human performance information to determine if predicted musculo response required by a machine operator is within the limits of capability for the machine designer's target operator population group. The exact same methods and computer programs that are used to create real-time man in the loop simulators for mechanical systems can thus be used by biomechanical groups to simulate human body response.

REAL-TIME OPERATOR IN LOOP SIMULATORS

MMIDAP intends to utilize operator in the loop simulators of complex mechanical systems to investigate environmental situations and work scenarios that have the potential for placing the machine operator in simulated harms way. These also provide a cost effective means of testing a variety of design options.

Major advances in formulating the mathematical equations needed to simulate complex mechanical equipment, along with the availability of low cost parallel processor computers have provided a unique opportunity to create low cost real-time simulators for complex mechanical equipment with the human operator in the control loop. Simulators accept real-time operator commands, predict system dynamic response, graphically create a simulated visual environment, and drive other laboratory devices to create a simulated vibrational and audio environment. Stress and load information for machine components can be obtained directly from the simulator. Qualitative control system feel and hand-eye coordination information can also be obtained from operator comments. Human performance data can be obtained by monitoring operator response in the simulated environment. The ability to simulate in real-time the gyrodynamic loads and machine force feedback control loads imposed upon the operator sets this new effort apart from that available in aircraft flight trainers.

The first step toward developing a simulator capability for complex mechanical systems was taken at the University of Iowa with its development of a simulator for the J. I. Case backhoe described in reference [13]. The ongoing second step is to create the Iowa Driving Simulator in 1991 is defined in reference [14] and the final step will be to develop the Department of Transportation's National Advanced Driving Simulator in the mid 1990's,defined in reference [15]. These capabilities can easily be modified to support the development of virtually any mechanical system to be operated by the intelligent physical exertions of a human operator.

MUSCLE MODELING AND LOAD SHARING

Detailed neuromusculoskeletal modeling of the human system or any of its subsystems is an extremely complex problem that is beyond today's state of the art capability. The First World Biomechanics Congress in August of 1990 had over 80 oral presentations on the subjects of multiple muscle systems, biomechanics, and movement organization. Formal reports on 46 of these presentations have been collected in reference [8]. From these reports and others presented at the Congress, it is clear that muscle dynamics and neuromusculoskeletal organization and movement modeling is a subject that will occupy researchers for many more years.

Complexities associated with modeling muscle contraction dynamics are matched by the problem of resolving the muscle load sharing and kinematic redundancy. The presence of redundant muscle actuators at virtually every anatomical joint implies that rules must exist for defining how muscles share the work load. An extensive summary of cost functions relevant to ongoing research in muscle load sharing at the University of Wisconsin at Madison has been provided in reference [9]. This work models the entire musculoskeletal system as a collection of hinged rigid bodies. Each muscle is modeled as a linear actuator that may wrap around bony structure and act along a resultant line of action between the points of muscle insertion and origin. To support this effort Williams, in references [16] and [17], has developed a computer program for determining muscle load sharing. The program uses Seireg's database of muscle characterization information. It computes the musculo force required by each muscle to maintain the system in quasi-static equilibrium. This work forms a major element of the work proposed in reference [5].

REPETITIVE WORK - FATIGUE, DISCOMFORT, AND INJURY

There is a good understanding of the causes of fatigue and repetitive motion injuries in industry, derived from understanding the physiology of muscle contraction and the treatment of many types of injuries. However, very little is known about how much exercise or work at any given level causes problems. Reference [5] outlines the process of validating modeling techniques associated with the investigation and prevention of industrial injury. This work will be done in collaboration with Wiker of the University of Wisconsin, director of the NASA sponsored Wisconsin Center for Space Automation and Robotics. This Center has a number of laboratories that are being made available to the MMIDAP project to conduct controlled experiments. For example, there are currently plans for Wiker and his students to investigate procedures for determining response relationships for muscle fatigue using muscle force calculations obtained from modeling software.

SUMMARY

The MMIDAP project can currently be viewed as three tightly coupled projects. Each of which can be justified independently. The union of the three however, will enable far greater capability than any of the parts. In summary:

o The four university MMIDAP team is addressing the fundamental problem of how to integrate a myriad of biomechanics and human performance related databases and analysis capabilities. Support for the disciplinary fields of sports medicine, orthopedics, prosthesis design, and neural functional stimulation are of major interest to researchers at the University of Iowa and Case Western Reserve University. Analysis needs within these fields of study will require enhancements of both existing biodynamics modeling and data collection capability. Support for the fields of rehabilitation engineering and physical therapy are of major interest at the University of Texas at Arlington. This will require the coupling of body response prediction with human performance analysis capabilities. The University of Colorado will be the provider of gross anatomy data. The matching of data analysis needs and data availability will define data gaps that can be filled as a by-product of the National Library of Medicine's Visible Human project. The details of creating an ability to associate human function with anatomy at a workstation is also being defined by the four university team.

o Photon Research Associates Inc. plans to develop the HMT-CAD (Human - Machine - Task Computer Aided Design) software system. Given a particular human body motion scenario required to achieve a particular work, recreation, or daily routine task, this software will assist in determining if a particular subject has sufficient physical resource capabilities (i.e. range of motion, strength, speed, endurance, etc.). This effort utilizes multibody dynamic analysis capabilities to predict human performance resource needs. It then couples these predictions to Kondraske's Basic Elements of Performance database at the University of Texas at Arlington to provide an estimate of task performance capability.

o R. J. Williams & Associates plans to develop the AEMUS (Analysis Engine for Musculoskeletal Systems) software system. An investigation of the potential of support from the market place has been quite revealing. While the research community is highly advanced in capability and need, the industrial market place is far less sophisticated. There is not a readily trained set of users prepared to use a capability designed by and for use in university research application. As a consequence, AEMUS must be developed in such a manner that it can grow with the sophistication of its market base. Its initial focus will be in support of industrial ergonomics and the investigation of repetitive work related problems of fatigue, discomfort, pain, and injury.

ACKNOWLEDGMENT

This project is supported by NASA's Small Business Innovative Research (SBIR) program. The continued support for this technology utilization effort by GSFC's Office of Commercial Programs, headed by Donald S. Friedman is appreciated by the author and all MMIDAP project collaborators.

REFERENCES

[1] - Frisch, H. P. "A Man/Machine Interaction Dynamics and Performance (MMIDAP) Analysis Capability," Second Annual Symposium on Mechanical System design in a Concurrent Engineering Environment, The University of Iowa, Oct 30-31, 1990.

[2] - Kroemer, K. H. E., et al. (Editors), "Ergonomic Models of Anthropometry, Human Biomechanics and Operator Equipment Interfaces," Proceedings, Workshop on Integrated Ergonomic Modeling, 1988.

[3] - G. V. Kondraske, J. G. Andrews, J. M. Mansour, V. Spitzer, D. Whitlock, and H. P. Frisch, "Plan for an Integrated Multiple Hierarchy Biomechanical Analysis and Modeling Tool," in preparation for submittal to CRC Critical Reviews in Bioengineering, contact H. P. Frisch, Code 714. 1, NASA/GSFC, Greenbelt, MD 20771, phone (301) 286-8730.

[4] - Turner, J. D. , "Integrated Ergonomic System Software Development," NASA/SBIR Final Report. J. D. Turner, Photon Research Associates, Inc., Cambridge Division, 1033 Mass. Ave., Cambridge MA 02133, phone (617) 354-1522.

[5] - WIlliams, R. J. , "Analysis of the Human Musculoskeletal System for Teleoperator System Design," NASA/SBIR Phase 1 Final Report. R. J. Williams & Associates, 631 Harriet Ave., Shoreview, MN 55126, phone (612) 483-0649.

[6] - "Electronic Imaging," Report of the Board of Regents, National Library of Medicine Long Range Plan, NIH Publication Number 90-2197, U. S. Department of Health and Human Services, April 1990.

[7] - Yamaguchi, G. T., Sawa, A. G. U., Moram, D. W., Fessler, M. J., and Winters, J. M., "A Survey of Human Musculotendon Actuator Parameters," Appendix of [8].

[8] - Winters, J. M. and Woo, S. L. (Editors), "Multiple Muscle Systems Biomechanics and Movement Organization," Springer Verlag, 1990.

[9] - Seireg, A. and Arvikar, R., "Biomechanical Analysis of the Musculoskeletal Structure for Medicine and Sports," Hemisphere Publishing Corporation, 1989.

[10] - Kondraske, G. V. "Quantitative Measurement and Assessment of Performance," Chapter 6 of book Rehabilitation Engineering, Smith R. V. and Leslie J. H. (Editors), CRC Press, 1990.

[11] - Kondraske, G. V., et al., "Measuring Human Performance: Concepts, Methods, and Applications," SOMA: Engineering for the Human Body (ASME), pp 6-13, January 1988.

[12] - Schiehlen, W. (Editor), "Multibody Systems Handbook," Springer Verlag, 1990.

[13] - Chang, J. L., Kim, S. S., and Haug, E. J., "Real-Time Operator in the Loop Simulation of Multibody Systems," Technical Report R-72, The University of Iowa, Center for Simulation and Design Optimization of

Mechanical Systems, 1990.

[14] - Stoner, J. W., et al., "Introduction to the Iowa Driving Simulator and Simulation Research Program," University of Iowa, Center for Simulation and Design Optimization of Mechanical Systems, Technical Report R-86.

[15] - Haug, E. J., et al., "Feasibility Study and Conceptual Design of a National Advanced Driving Simulator," US Department of Transportation, DOT HS-807-596, 1990.

[16] - Williams, R. J. and Seireg, A., "Interactive Computer Modeling of the Musculoskeletal System," IEEE Trans on Biomedical Engineering, BME-24, pp 213-219, 1977.

[17] - Williams, R. J. and Seireg, A., "Interactive Modeling and Analysis of Open or Closed Loop Dynamic Systems with Redundant Actuators," Journal of Mechanical Design, Vol 101, pp 407- 416, 1979.

SOFTWARE ENGINEERING

(Session B6/Room B1)

Wednesday December 4, 1991

- Hybrid Automated Reliability Predictor Integrated Workstation (HiREL)
- Using Ada and the Rapid Development Lifecycle
- Advances in Knowledge-Based Software Engineering
- Reducing the Complexity of Software Development Through Object-Oriented Design

N92-22468

HYBRID AUTOMATED RELIABILITY PREDICTOR INTEGRATED WORK STATION

HIREL

Salvatore J. Bavuso

NASA Langley Research Center

Hampton, VA 23665-5225

ABSTRACT

The Hybrid Automated Reliability Predictor (HARP) integrated reliability (HiREL) work station tool system marks another accomplishment toward the goal of producing a totally integrated computer-aided design (CAD) work station design capability. Since a reliability engineer must generally graphically represent a reliability model before he can solve it, the use of a graphical input description language increases productivity and decreases the incidence of error. The captured image displayed on a cathode ray tube (CRT) screen serves as a documented copy of the model (as a hard copy can be readily made by the push of a button) and provides the data for automatic input to the HARP reliability model solver. The introduction of dependency gates to a fault tree notation allows the modeling of very large fault tolerant system models using a concise and visually recognizable and familiar graphical language. In addition to aiding in the validation of the reliability model, the concise graphical representation presents company management, regulatory agencies, and company customers a means of expressing a complex model that is readily understandable. The graphical postprocessor computer program HARPO (HARP Output) makes it possible for reliability engineers to quickly analyze huge amounts of reliability/availability data to observe trends due to exploratory design changes. HiREL is written in ANSI standard code for maximum portability and has been successfully executed on IBM compatible 286/386/486 personal computers, Sun and Vaxstation platforms. The major components of HiREL have already proven themselves to be a useful modeling asset to a number of aerospace companies that have been serving as beta test sites since 1985.

INTRODUCTION

Electronic design engineers are increasingly faced with shorter design cycle times which account to a large extent for the heightened interest in computer-aided design software (CAD) tools. In conjunction with the advent of affordable powerful work stations, CAD software is becoming a mainstay capability in the engineering community. The success of current CAD tools has encouraged the engineering community to seek a capability that totally integrates the system design process. Although this capability does not presently exist, many of the software components that are required for such a capability are presently available.

One such software component that has recently been developed and released to the engineering community is HiRel: the Hybrid Automated Reliability Predictor (HARP) integrated Reliability system tool for reliability/availability prediction (ref. 1). HiRel offers a toolbox of integrated software modules that can be used to customize the user's application in a work station environment. It consists of two interactive graphical input/output modules and four reliability/availabilty modeling engines that provide analytical and simulative solutions to a wide host of highly reliable fault-tolerant system architectures and is also applicable to electronic systems in general.

The tool system was designed at the outset to be compatible with most computing platforms and operating systems and some modules have been beta tested within the aerospace community for over seven years. Over 100 copies have been distributed. Many examples of its use have been reported in the literature and at the HARP Workshop conducted at Duke University, July 10-11, 1990 (ref. 2).

HIREL DEVELOPMENT

The development of HiRel has been an evolutionary project that has spanned over nearly two decades. The goal that was set for HiRel circa 1973 was to develop a capability to assess the reliability/availability of any fault-tolerant digital computer-based system, including the system effects of software, i.e., fault/error handling. Although the initial target application was for assessing highly reliable real-time digital fault-tolerant aircraft flight control systems, the developers of HiRel did not limit its applicability to soley that application. The realization that one day NASA spacecraft would require ultrahigh fault-tolerant systems, motivated the HiRel developmental team to include a reliability modeling capability to accurately represent non-constant failure rate models as well as constant failure rate models typically found in aircraft applications. Data to support the use of the decreasing failure rate model were published as early as 1975 and provided much of the motivation (ref. 3). This foresight was fortuitous since there now exists significant data to productively use HiRel's Weibull decreasing failure rate parts model. For very long manned missions, such as a mission to Mars presently under consideration by NASA, decreasing failure rate models may well prove to be the modeling technology that can provide the reliability confidence to make such a trip.

HIREL DESCRIPTION

The wide range of applications of interest has caused HiRel to evolve into a family of independent software modules that communicate with each other through files that each module generates. In this sense, HiRel offers a tool box of integrated software modules that can be executed to customize the user's application. Figure 1 depicts the HiRel tool system. The core of this capability consists of the reliability/availability modeling engines, which are collectively called the Hybrid Automated Reliability Predictor (HARP). It is comprised of four self-contained executable software components: The original HARP module, Monte Carlo HARP (MC-HARP), Phased Mission HARP (PM-HARP), and X -Windows HARP (XHARP). In conjunction with the engine suite, there are two input/output interactive Graphical User Interface (GUI) modules that provide a work station environment for HiRel. These software modules are called the Graphics Oriented (GO) module and the HARP Output (HARPO) module.

Reliability/Availability Modeling Engines

The power of HiRel resides in its engine suite which consumed the bulk of the development effort and took over a decade to complete. A mathematical modeling technique called behavioral decomposition and a fault tree notation called dynamic fault trees constitute the major engine suite modeling power which is used by the other engine suite modules. The engine suite is composed of independently executable software modules: HARP/behavioral decomposition, HARP/dynamic fault tree model, MC-HARP, PM-HARP, and X-HARP. Since the original HARP is the kernel of the engine suite, it will be discussed first.

HARP/Behavioral Decomposition Modeling Engine

The prototype reliability/availability engine was implemented into the "Textual" HARP software subsystem which initially was a stand-alone system that later became integrated into HiRel. Textual HARP offers a textual interactive input capability when executed in a stand-alone mode and will execute on many computing platforms requiring only an ANSI standard FORTRAN compiler. The GO software module is offered as a complementary input capability to the textual input format but requires the installation of a commercially available graphical software package for execution.

HARP is a software tool for analytically predicting the reliability/availability of fault-tolerant digital computer-based systems; however, it is also applicable to a very large class of systems in general. In addition to reliability/availability, it can be used to analyze system sensitivity and failure causes. Its notable features include: very large system modeling, dynamic fault modeling, automatic conversion of fault tree input to a Markov chain or manual Markov chain input, automatic insertion of fault handling models into Markov chains, automatic parametric analysis, and wide portability of the code. It utilizes a method called behavioral decomposition to solve for the reliability of a system when fault/error handling is modeled. A discussion on this subject follows; however, the reader should see references 4 and 5 for more details.

When fault/error handling is considered, dependencies exist between stochastic events that make it necessary and practical to use a Markovian representation of the reliability model. A Markov process contains information about a system's fault processes, component depletion, and recovery procedures. Graphically, a Markov model consists of states and transitions. The states contain information about the number of operational components, and the transitions are rates at which specific components or subsystems fail causing a change in the state of the system. Computations are done to determine the probability of being in a state based on time. The reliability of the system can then be determined by adding the probabilities of the operational states (ref. 4). However, in systems designed with fault-tolerance, a very large state space model can result which introduces computational problems. These problems can be solved by utilizing the methods of decomposition and aggregation, i.e., dividing the system into smaller subsystems based on component types, solving these models separately, and then combining the results of the subsystem models to produce the larger system's solution. However, this method requires that the behaviors of the subsystems be independent. In many fault-tolerant systems this is a false assumption, because these systems may include dependencies. HARP, however, offers a unique modeling technique that surmounts this potential difficulty.

In addition to this traditional modeling technique, HARP offers a simpler and more efficient approach called behavioral decomposition (ref. 6). Using this method, HARP allows a user to segregate a reliability/availability model into two submodels, a fault-occurrence/repair model (FORM) and fault/error handling model (FEHM). The FORM describes a system as a fault tree or a Markov chain and relates information about hardware redundancy and fault processes. Using a FEHM to describe specific recovery procedures, a user can include details about permanent, transient, and intermittent faults in a reliability model. Figure 2 illustrates the behavioral decomposition method utilizing FORM and FEHM submodels. HARP provides a user with seven FEHMs which range from a simple probabilities and moments FEHM to a very complex extended stochastic Petri net FEHM. The model can be input into HARP by using an interactive textually oriented interface or a graphically oriented interface. If the FORM is a fault tree, it is first converted to a complex stochastic process that is reduced to a simpler Markov chain. The FEHMs are solved separately from the FORM to determine the exit probabilities and holding times for transient restoration, permanent coverage, near-coincident fault failures, and single-point failures. No matter how complex the FEHM models may be and no matter how many FEHMs are specified, this process will produce at most two additional system failure states in the chain which represent near-coincident fault failures and single-point failures. The reduction of an enormous number of Markov states for most practical systems is the forte of behavioral decomposition. The model is then given to a popular ordinary differential equation solver to compute the results.

HARP/Dynamic Fault Tree Modeling Engine

Input to HiRel takes one of two forms that can be either specified textually or graphically. In either case, the user can specify a FORM in the Markov graph or fault tree notation (ref. 7). The standard input to HiRel is the fault tree notation which consists of the standard combinatorial gates, AND, OR , and M out of N. Four special fault tree gates that allow sequence and functional dependencies have been added to provide a dynamic FORM modeling capability. The notational simplicity and power of these dynamic gates is demonstrated in references 8 and 9.

The functional dependency gate is depicted in figure 3. It is the logical equivalent of a combinatorial fault tree composed of AND and OR gates when no fault handling is specified. The input labeled "trigger" can be the output from any gate, whereas the outputs take two forms. The non-dependent output simply mimics the trigger input and may or may not be connected to any input of any gate, i.e., it can dangle if desired. The typical use of this gate involves the other outputs. The outputs labeled "dependent events" must be basic events. Although they are labeled dependent events, the basic events themselves are statistically independent. The dependency is related to the trigger event. A typical use of this gate is to account for the functional loss of devices because some other device failed and therefore is unable to provide signal or power input to the downstream operational devices.

A non-combinatorial gate that implements a cold spare model appears in figure 4. This sequence dependency gate is naturally called the cold spare (CSP) gate. The gate output fires (produces an output) when and only when the primary event occurs first followed by events 1st, 2nd, ..., nth. Events 1st, 2nd, ..., nth cannot occur first. Thus, the primary event can represent an active unit and event 1st is the cold spare that exhibits a zero failure rate until the active unit fails. At that instant, the cold spare is powered up and immediately exhibits a failure rate greater than zero. If additional cold spare units are added, they are powered up in the order of left to right and all inputs are independent basic events.

A useful variation of the CSP gate is called the sequence enforcing gate (fig. 5). The inputs of the sequence enforcing gate can be basic events or the output of some other gate for the primary input only. The sequencing of events is left to right similarly to the CSP gate. The cold spare gate and the sequence enforcing gate differ primarily in the way they treat shared events.

The last dynamic sequence dependency gate is called the priority AND (P-AND) gate (fig. 6). The P-AND gate differs from a combinatorial AND gate in only one respect: In HARP, only two inputs for the P-AND are allowed, and the gate produces an output only if the left most event occurs first followed by the right most event. Contrary to the CSP gate, the right most event in a P-AND can occur first, but no output results. The functionality, the name of the gate, and the gate symbol were obtained directly from the literature (ref. 10).

The developers experience with the use of these new gate additions to HARP has been extensive. They have applied them to some very complicated fault-tolerant network systems (ref. 8). Although there is no warm spare gate, that model has been functionally emulated with the existing gates, and pooled spares models have also been emulated. With HARP's Markov chain truncation technique that bounds the truncation error, extremely large Markov chains have been modeled and solved that have simple appearing fault tree diagrams. These models have demonstrated the succinct yet powerful notational value of HARP's dynamic fault tree capability.

An additional gate, the NOT gate, was added to HARP for completeness but was commented-out in the source code because its inclusion allows the modeling of noncoherent models. A noncoherent model allows the possibility of the top event of a fault tree to exhibit a decreasing probability of failure with increasing time. The HARP Team wanted to properly document the use of the NOT gate because of the likelihood of misuse. That documentation has not yet been completed; however, researchers at Duke University mathematically proved that the complete set of HARP's fault tree gates maps into the set of non-cyclic Markovian models with constant transition rates (ref. 9). Although there are no plans to further extend this capability at Langley, the most obvious and useful further extension should include a fault tree notation to model repair.

MC-HARP Modeling Engine

Simulation for use in reliability prediction has been used for decades. The traditional simulation method is called analog Monte Carlo which relies on a large number of failure events occurring in the mission time of interest. In highly reliable systems, the first system failure event occurring at time t is not likely to occur until $t > T$, where T is the mission time. Thus an inordinate number of simulation trials are required to produce an acceptable confidence level.

Recently, variance reduction techniques called importance sampling have been rediscovered by the reliability community. Importance sampling is a technique that was reported in the literature as early as 1984 (ref. 11). The basic concept of importance sampling is to force and bias transitions along the rare event paths in an underlying Markovian model (which may contain both local and global time dependence with a disparity of typically six orders of magnitude) while dynamically maintaining a record of the forcing and biasing that allows post simulation construction of an unbiased estimator of the event of interest, (e.g., system failure) with extremely low variance. The prime challenge over the years has been one of determining a suitable failure biasing scheme.

MC-HARP was developed to model non-Markovian models which arise when systems exhibit nonconstant failure rate histories and when cold or warm spares are employed (ref. 12). This failure history is a possible scenario for systems to be used in manned deep-space missions. The other motivation was to developed a modeling capability for correlated transient induced failures as might occur when an aircraft system is exposed to high intensity radio frequency emissions, e.g., lightning. Preliminary applications of MC-HARP as reported in reference 12 are very encouraging. Several highly reliable systems were analyzed and compared to the HARP analytical results. For large systems, MC-HARP proved more efficient particularly for non-constant failure rate models. MC-HARP can also be used to circumvent behavioral decomposition to serve as a check or to replace it.

PM-HARP Model Engine

Phase Mission HARP was developed by the University of Washington for Boeing Electronics and Aerospace Company (ref. 13). A phase is an epoch in a mission. During an epoch, a system may be altered by external means. An example of a phase occurs when the failure rates of the initial system change perhaps because of some environmental stress. A spacecraft system during launch would experience more vibration and shock than during orbital operation which would be a second and more benign phase. Another example of a phase mission occurs when a system is tested and repaired prior to the continuation of service of a commercial aircraft. During testing and repair, the system may not have been fully restored perhaps due to imperfect diagnostics or repair. The phase time may be deterministic or stochastic. PM-HARP was developed to facilitate this class of modeling and analysis.

XHARP Modeling Engine

Aside from the desirable portability of X-Windows HARP implemented in the X-Windows environment (X-HARP), X-HARP offers a unique automatic behavioral decomposition capability that was never implemented in the original HARP (ref. 14). The new fault/error handling modeling capability was developed to assist users who are unsure of the specifics of using the standard behavioral decomposition model. X-HARP further extends the multi-fault modeling capability of the original HARP to allow multiple entry and exit transitions to user specified fault/error handling models. Also, X-HARP allows the user to specify a detailed multi-fault model for system designs that use fault containment regions. Although the original HARP multi-fault models which were designed to be easy to use and specify will produce a conservative pessimistic unreliability prediction, for some system designs such as those with fault containment regions, the original HARP model may produce an overly conservative result. When an overly conservative result is unsatisfactory, X-HARP in conjunction with HARP, will produce a more accurate prediction commensurate with the accuracy of the user's data.

HIREL - Interactive Graphical User Interface

Graphics Oriented Interactive Input

The graphics oriented (GO) module enables the user to "draw" reliability models in the form of fault trees or Markov chains on the screen of a work station monitor (ref. 1). Figure 7 depicts the screen image for the fault tree drawing mode. A click of the "mouse" button device toggles the display to show the dynamic fault tree gates as shown in figure 8. A gate is drawn by selecting the draw primitive followed by selecting the particular gate to be drawn. Using the mouse, the cursor is positioned in the drawing screen area to the left, and the gate is moved to that position. Subsequent gates to be drawn are simply selected as required. The labeled squares on the right hand side of the screen in figures 7 and 8 are also selected with a cursor under control of a moving mouse device . The functions they provide are evident as labeled. Figure 9 displays the screen image for the Markov chain drawing mode. The drawing primitive provided by the four rectangles on the lower right are selected with a cursor as described previously to draw the chain model.

The captured image displayed on the screen serves as a documented copy of the model, as a hard copy can be readily made by the push of a button. The data from which the image was created also serves as the data for automatic input to the HARP reliability model solver. The introduction of dependency gates to a fault tree notation which provides a dynamic fault tree capability, allows the modeling of very large fault tolerant system models using a concise and visually recognizable and familiar graphical language. In addition to aiding in the validation of the reliability model, the concise graphical representation presents company management, regulatory agencies, and company customers a means of expressing a complex model that is readily understandable.

HARP Output

The graphical postprocessor module HARPO (HARP Output) makes it possible for reliability engineers to quickly analyze huge amounts of reliability/availability data to observe trends due to exploratory design changes (ref. 1). HARPO reads files automatically generated by the HARP modeling engine. It will accept files from any previously generated HARP executions for comparative analysis. The user can in an interactive mode display up to nine graphs representing modeling iterations of the same system or compare different system models. A number of parameters can be altered for analysis such as failure rate or coverage data for sensitivity analysis, or to view the effects of unreliability/unavailability as a function of different mission times. Markovian state probabilities or sums of user specified state probabilities can also be displayed. The user can manipulate HARP and HARPO ASCII files to do performability computations and display them. HARPO uses the ANSI standard Graphics Kernel System (GKS) graphics software which allows portability and provides a large number of device drivers to output graphical data to many hard copy devices, e.g., laser printers and plotters.

Figure 10 depicts a typical screen image for HARPO. The graph shows the probability of system failure versus time for two parameters of interest (M3F2 and M3REXHST) for the 5th version of model number 3 for a two processor - two bus system (MODEL3 5 3p2b). M3F2 designates "failure state number two" for model number three, where F2 is a failure state of the given Markov chain. M3REXHST designates the sum of all the failure states (including F2) that caused system failure resulting from the exhaustion of redundant hardware modules. In fault-tolerant systems, system failure can also be caused by improper fault/error handling not shown in this graph. The title RS - STATES tells the reader the data came from the RS (results) file generated by the HARPENG module.

HIREL PORTABILITY/AVAILABILITY

HARP was developed on a Sun 3 computing platform running under Berkley Unix 4.3. The source code was written in ANSI standard FORTRAN 77. HARP has been ported to a large host of computing platforms with the major operating systems being DEC VMS and Ultrix, Berkley Unix 4.3, AT&T Unix 5.2, and MS DOS. PC-HARP running under MS DOS is a scaled down version of HARP that executes on IBM compatible 286/386/486 machines. Certain limitations are placed on PC-HARP's capabilities because of the 640 K memory restriction imposed by MS DOS; however, extended FORTRAN compilers such as Lahey F77l-Em/16 and their DOS Extender are commercially available which use extended memory and thus breaches the 640K memory MS DOS barrier. Full scale HARP code can be compiled and executed without modeling restrictions on 286/386/486 PC compatible machines with extended memory. Operating systems other than MS DOS such as OS/2 and Unix, also allow full scale HARP to successfully execute with sufficient extended memory. Most of the HiRel software modules are available through NASA's software distribution facility, COSMIC[1] or from the developers at Duke University[2]. The MC-HARP[3], PM-HARP[4], and the XHARP[5] Engines are available from the universities where they were developed.

---

[1] COSMIC, The University of Georgia, 382 East Broad St., Athens, GA 30602 (404) 542-3265

[2] Duke University, Dept. of Computer Science, Durham, NC, 27706, Joanne B. Dugan, (919) 660-6559

[3] Northwestern University, Dept. Mechanical Eng., Evanston, IL 60206, E. E. lewis (708) 491-3579

[4] University of Washington, Dept. of E. E., Seattle, WA 98195, Arun. K. Somani (206) 685-1602

[5] Clemson University, Dept. of Computer Sci., Clemson, SC 29734-1906, Robert Geist (803) 656-2258

HIREL APPLICATIONS -

The core HiRel module, HARP, has been applied to numerous applications in the seven years of beta testing. Some of these applications are listed as follows:

Aircraft Life Critical Systems, Civilian Aircraft Electronics, Military Avionics, Space Systems, Computer Systems, Railroad Control Systems, Nuclear Power Control Systems, Submarine Steering Control Systems

For more detail on specific systems and architectures where HARP has been applied, see references 5, 7, 9, 10, 12, and 13. Also see the proceedings of the Duke/HARP Workshop (ref. 2).

REFERENCES

1. Salvatore J. Bavuso and Joanne Bechta Dugan: HiRel - Reliability/Availability Integrated Work Station Tool, Reliability and Maintainability Symposium Proceedings, Jan. 22 - 24 ,1992

2. K. S. Trivedi and S. J. Bavuso, compilers: HARP Workshop, Duke, University, July, 1990.

3. A. R. Timming: A Study of Total Space Life Performance of GSFC Spacecraft, NASA TN D-8017, 1975.

4. Joanne B. Dugan, Kishor S. Trivedi, Mark K. Smotherman, Robert M. Geist: The Hybrid Automated Reliability Predictor, Journal of Guidance, Control, and Dynamics, Vol. 9, No. 3, May-June 1986.

5. Salvatore J. Bavuso, Joanne B. Dugan, Kishor S. Trivedi, Elizabeth M. Rothmann, and Mark A. Boyd: Applications of the Hybrid Automated Reliability Predictor, NASA TP 2760, December 1988.

6. Robert Geist, Mark Smotherman, Kishor Trivedi, and Joanne Bechta Dugan: The Reliability of Life Critical Systems, Acta Informatica 23, 621-642 (1986).

7. Elizabeth M. Rothmann, Joanne Bechta Dugan, Kishor S. Trivedi, Mark A. Boyd, Nitin Mittal, Salvatore J. Bavuso: HARP: The Hybrid Automated Reliability Predictor Introduction and Guide for Users, Version 6.1, Dept. of Computer Science, Duke University, May 1990.

8. Joanne Bechta Dugan, S. J. Bavuso, and M. A. Boyd: Modeling Advanced Fault-Tolerant Systems with HARP, Tutorial Presented at the Annual Reliability and Maintainability Symposium, January, 1991.

9. Mark A. Boyd: Dynamic Fault Tree Models: Techniques for Analysis of Advanced Fault Tolerant Computer Systems, Phd Dissertation, Dept. of Computer Science , Duke University, April, 1991.

10. J. B. Fussell, E. F. Aber, and R. G. Rahl: On the Quantitative Analysis of the Priority-AND Failure Logic, IEEE Transactions of Reliability, R-25(5), 324-326, December 1976.

11. E. Lewis and F. Bohem: Monte Carlo Simulation of Markov Unreliability Models, Nuclear Engineering and Design, 77 (1984).

12. M.E. Platt, E. E. Lewis, and F. Bohem: General Monte Carlo Reliability Simulation Code Including Common Mode Failures and HARP Fault/Error Handling, Dept. of Mechanical Engineering, Northwestern, University, Final Report for NASA Grant NAG-1-1031, January 1991.

13. Arun K. Somani, James A. Ritcey, Stephen H.L. Au, and Hamid Amindavar: Phased Mission Analysis Program Users Manual, Dept. of Electrical Engineering, University of Washington, December 1989.

14. Robert Geist: Extended Behavioral Decomposition for Estimating Ultrahigh Reliability, IEEE Transactions on Reliability, Vol. 40, NO. 1, April, 1991.

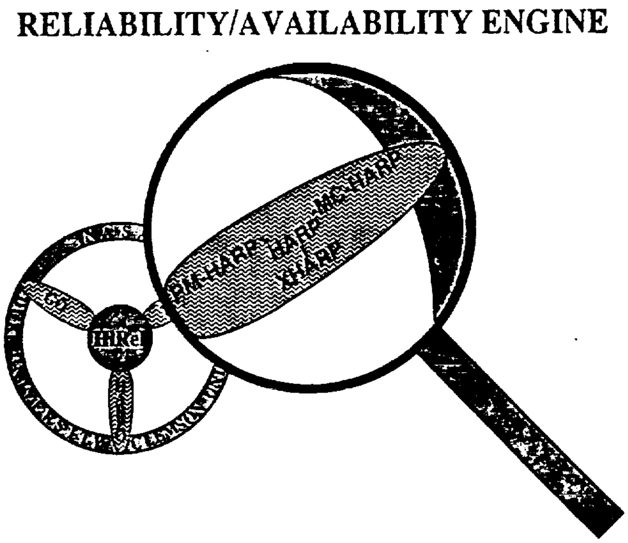


Figure 1: HiRel tool system depicting the modeling engine suite

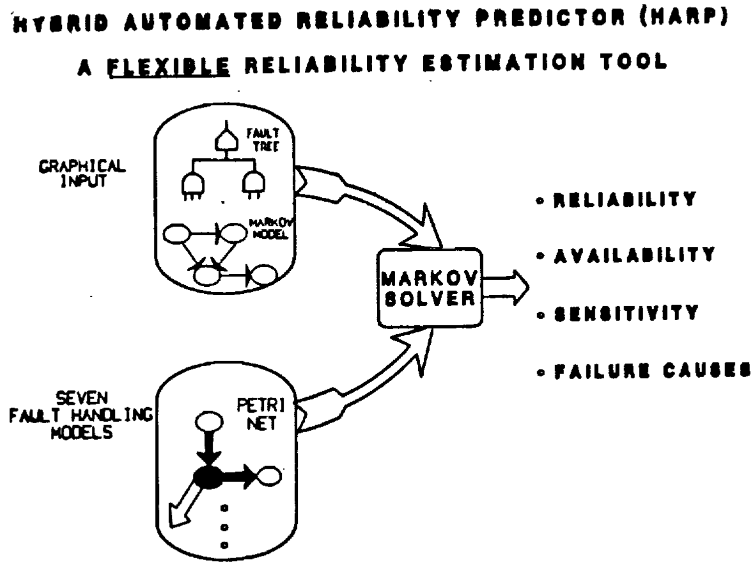


Figure 2: HARP/behavioral decomposition

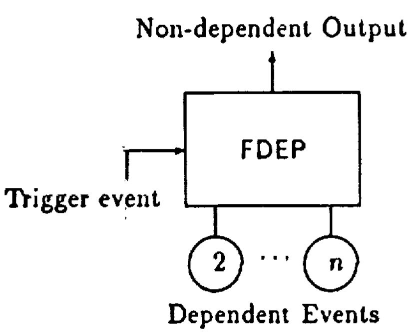


Figure 3: Functional dependency gate

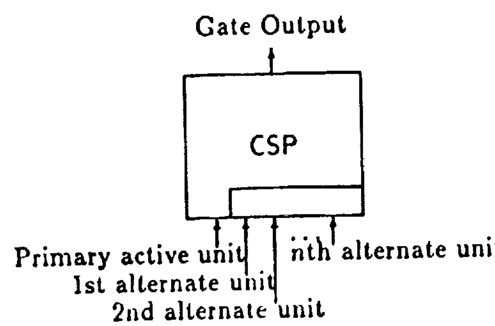


Figure 4: Cold spare gate

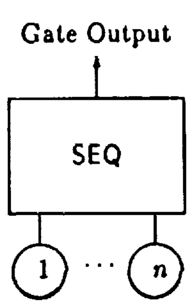


Figure 5: Sequence enforcing gate

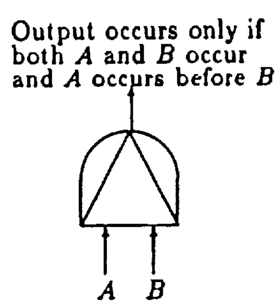


Figure 6: Priority-AND gate

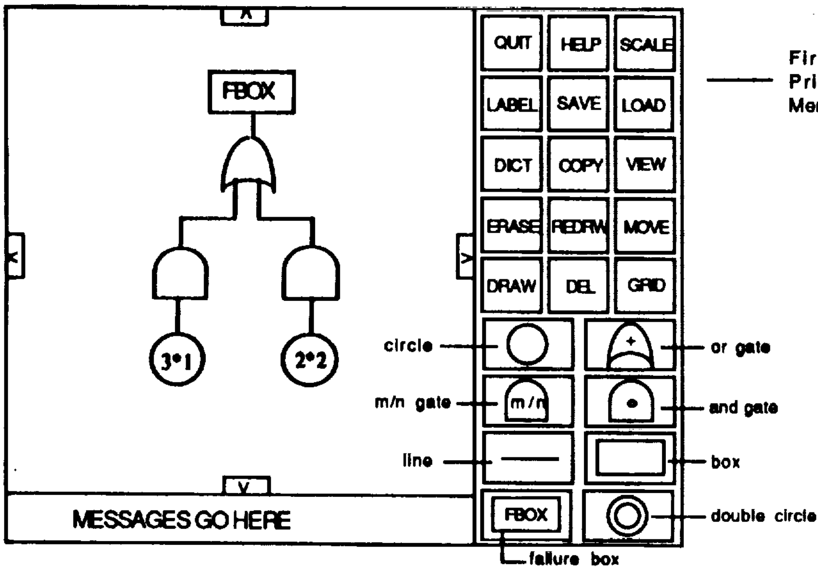


Figure 7: GO Module Screen Image
for a Fault tree model

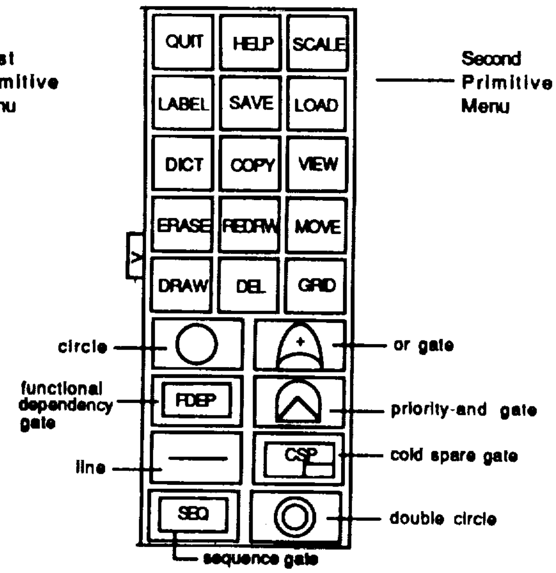


Figure 8. Primitive Menu for Sequence
Dependency Gates

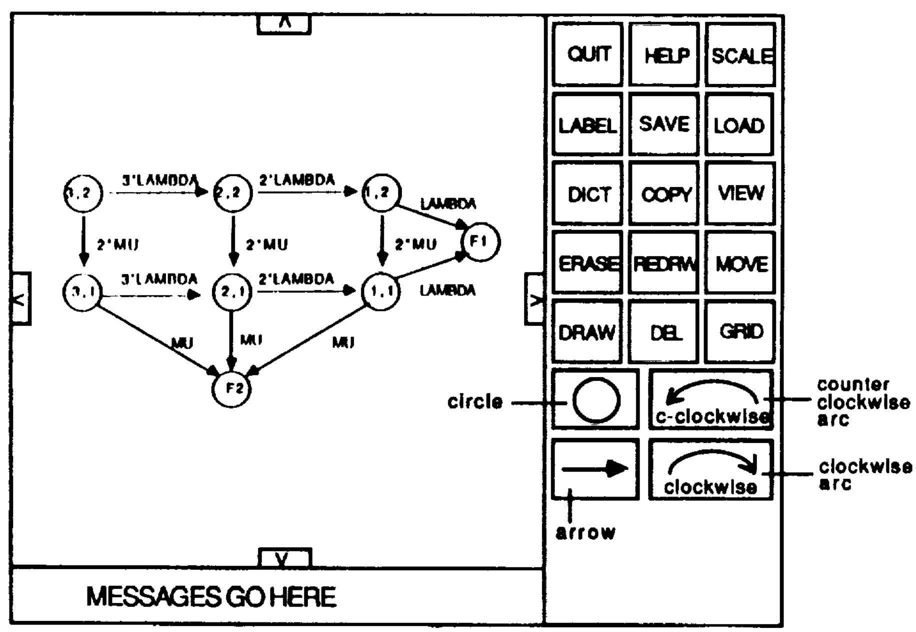


Figure 9. GO Module Screen Image
for a Markov Chain Model

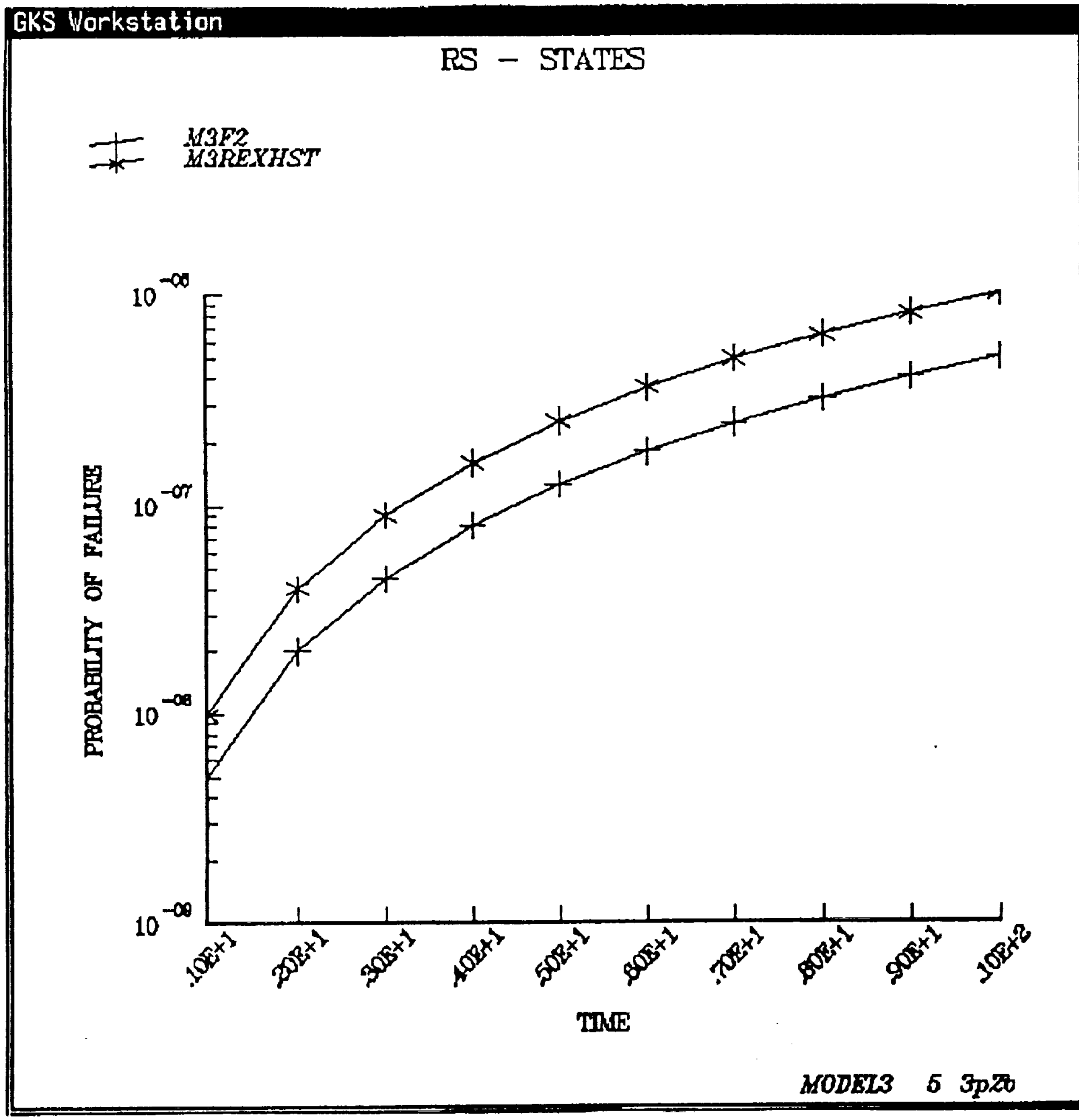


Figure 10. HARPO Module Screen Image for a 3-processor, 2-Bus System

N92-22469

Ada AND THE RAPID DEVELOPMENT LIFECYCLE

Lloyd DeForrest
Dr. Lynn Gref
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91009

ABSTRACT

JPL is under contract, through NASA, with the U.S. Army to develop a state-of-the-art Command Center System for the United States European Command (USEUCOM). The Command Center System will receive, process and integrate force status information from various sources and provide this integrated information to staff officers and decision makers in a format designed to enhance user comprehension and utility. The system is based on distributed workstation class microcomputers, VAX- and SUN-based data servers, and interfaces to existing military mainframe systems and communication networks.

JPL is developing the Command Center System utilizing an incremental delivery methodology called the Rapid Development Methodology with adherence to government and industry standards including the UNIX operating system, X Windows, OSF/Motif and the Ada programming language. Through a combination of software engineering techniques specific to the Ada programming language and the Rapid Development Approach, JPL has been able to deliver capability to the military user incrementally, with comparable quality and improved economies of projects developed under more traditional software intensive system implementation methodologies.

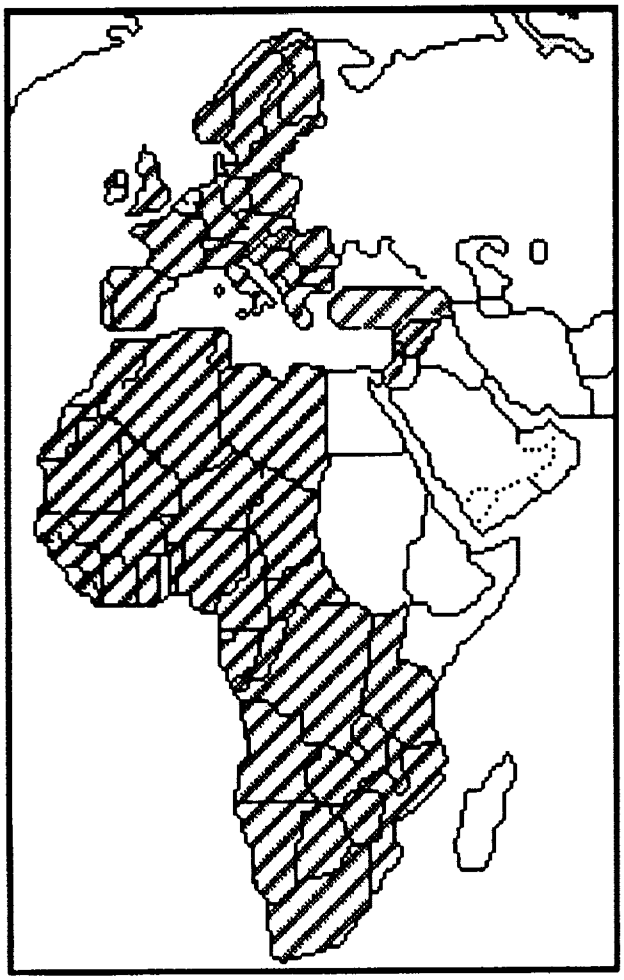
INTRODUCTION

The objective of the EUCOM Command Center Project (ECCP) is to provide USEUCOM a state-of-the-art Command Center System which integrates information available from approximately one hundred communications and automation systems. JPL is designing, implementing, installing, and testing this "core" system. The system will serve the Command staff in their mission execution. Figure 1 illustrates USEUCOM's area of responsibility and its basic mission. The area of responsibility covers all of western Europe and includes most of Africa and the Mediterranean Sea to the Suez Canal. The Command has responsibility for all unilateral actions of the U.S. military within its area of responsibility. Many world events of the past two years have involved USEUCOM. Disintegration of the Soviet Block, the Gulf War, the hostage releases in Lebanon and the U.S. Kurdish relief effort are just a few of these events affecting the Command. As a consequence, with each new crisis action, new requirements are levied on the Command Center System to process and present force status information in new or different ways. The impact of these new requirements is compounded by a shrinking budget for implementation of the Command Center System that has resulted in the demand to do more for less.

The following general functional capabilities have been identified as the minimum set that the Command Center System requires:

1. Available information shall be passed, processed and/or aggregated in a timely manner.
2. Information shall be automatically distributed, plotted and correlated.
3. Status and location of all forces shall be consolidated, integrated, and presented.
4. Presentation of available information shall facilitate the decision making process.
5. Integrated access to data from multiple security levels will be made available at the earliest opportunity.
6. Communications facilities shall support secure automatic voice, data and teleconferencing.

The overall architecture for the core system which will provide the above cited functional requirements is depicted in Figure 2. This architecture recognizes the need for an integrated approach for the three mediums of information: visual (video), voice, and digital data. A communications network for each security level and each medium



- Unified Command
- Area of Responsibility (AOR)
  - Europe, Africa, Mediterranean
  - 13 Million Square Miles
  - 350,000 U. S. Troops
- Strategic and Operational Tasks
  - Conventional, Nuclear, Special
- Continuus Operations
  - Peacetime
  - Crisis Management
  - Warfare
- Post-Conventional Forces Europe Functions
  - Monitoring / Directing Out-Of-Area Operations
  - Monitoring Compliance With Force Reduction Treaties
  - Coordinating Expanded Intel And Recon Services
  - Arranging / Directing Rapid Force Reduction

Figure 1. U.S. European Command Area of Responsibility and Mission

interconnects the various offices (i.e. sections) of the Command. Data Local Area Networks (LAN) of different security classification levels are interconnected by specialized security gateways. These gateways initially provide data transfers in one direction, from a lower security classification to a higher security classification. Future generations of the security gateways will provide limited two way data transfers.

The U.S. Army has adopted a number of standards for the implementation of Command Center Systems. These standards are designed to promote the sharing and reuse of software among Command Center Systems. Figure 3 depicts the layered software architecture with the associated standards being followed by the Project. In addition, the Project's documentation follow that required by the DoD-STD-2167A [7]. The adherence to these standards have presented many technical challenges with regards to software development. For example, bindings between the Ada programming language and the X Window/MOTIF user interface had to be developed. This software required a moderate effort to develop (approximately 15,000 lines of code and two work years) and has been provided to various military and commercial organizations with military sponsor approval (we are now in the process of placing this software into COSMIC). Another example is the development of a set of Application Support Layer libraries which facilitate the current and future development of user interface, map access, decision support and monitor and control applications.

Further challenges to the software development within the ECCP are: 1) a constrained and varying funding profile, 2) adherence to newly adopted standards, 3) utilization of Commercial Off-The-Shelf (COTS) and Non-Developmental (NDI) software whenever possible, 4) development of software in Ada, 5) employment of incremental deliveries every 9 to 12 months, and 6) the rapid turnover of user community members (users involved with the requirements definition and design are frequently not the actual users of the operational system).

The remainder of this paper provides a description of our approach to solving these challenges through a discussion of the Rapid Development Methodology, a brief discussion of the Ada programming language (given as background information), a discussion of the ECCP's combination of Ada and the Rapid Development Methodology, a discussion of the benefits of this approach (with some pitfalls), and a discussion of some of the lessons we have learned that might be beneficial to others.

RAPID DEVELOPMENT METHODOLOGY

The challenges briefly described in the previous paragraphs necessitate an unconventional development methodology. It has been our experience that the conventional wisdom upon which the classic "waterfall" approach [1] depends requires the following conditions be met, or at least maximized, for a successful project development:

1. Requirements be identified and configuration managed early in the development cycle [ref. 1, pg. 39]. This requires that the user articulate program behavior requirements sufficiently before the design phases.
2. A budget and funding profile be established up-front and followed, minimizing perturbations in Project team staffing.
3. Users can forgo getting the new capabilities for the 3 to 8 year development cycle.
4. The systems on the other side of external interfaces will not change unexpectedly during the development.

The Incremental Waterfall [1], and the Spiral Model [2,3,4] methodologies are two examples of methodologies that address, and attempt to minimize, the effects of change to the software development induced by the above conditions not being met. Additionally, JPL has developed a software engineering methodology similar to the Waterfall and Incremental Waterfall methodologies called the "JPL Software Management Standards" [5].

The Rapid Development Methodology (RDM) is similar to the Incremental Waterfall, Spiral and institutional JPL methodologies, but has been tailored to specifically address the above mentioned conditions. RDM has been used successfully on previous JPL command center improvement projects (e.g. the Distributed Management Information and Communication System {DMICS} and the Global Decision Support System {GDSS}).

Another variation of the incremental development methodologies is the incremental delivery of prototypes (rapid prototyping) which could be used (as was used during the early stages of the DMICS) to continually establish and

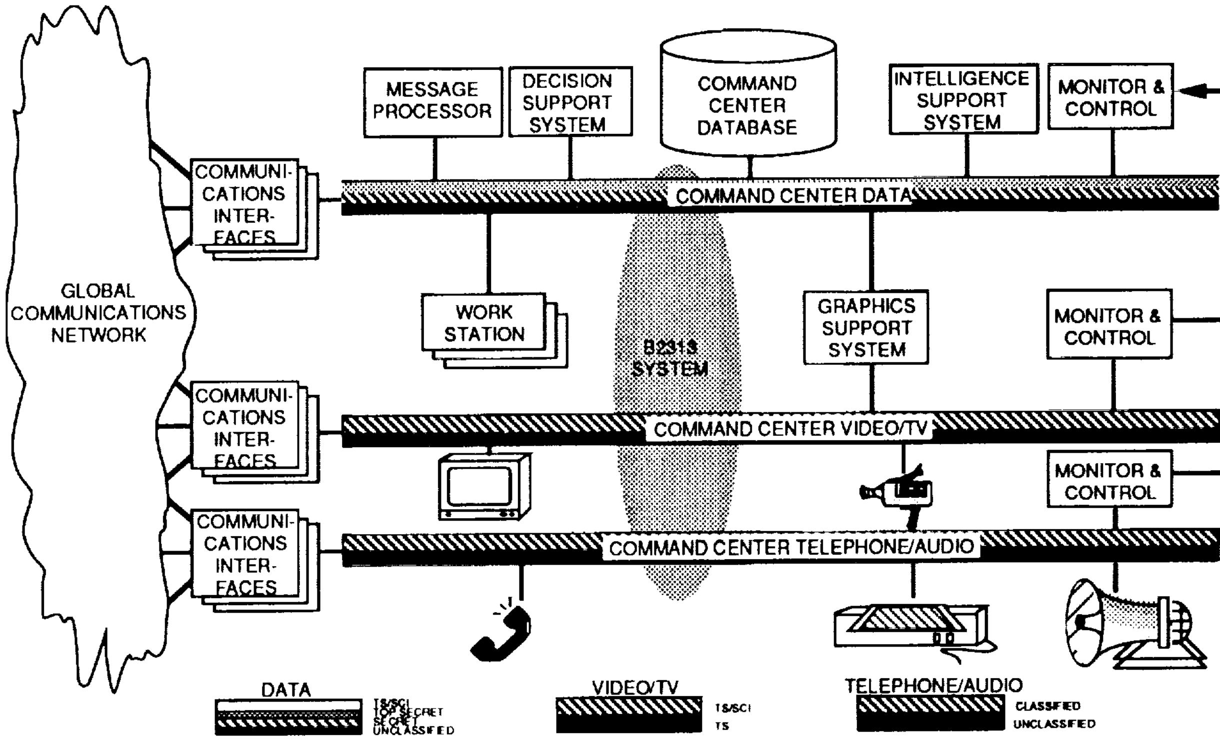


Figure 2. Generic Command Center Architecture

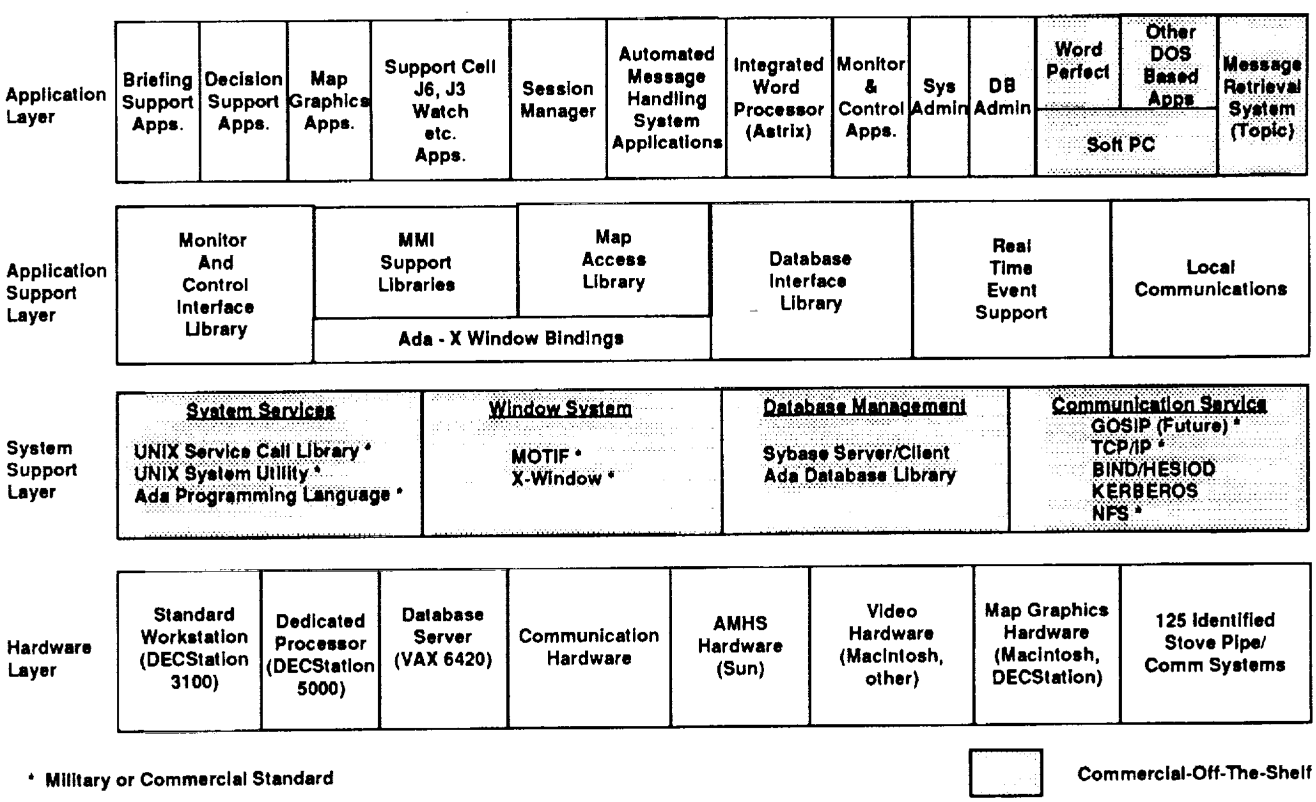


Figure 3. EUCOM Project Layered Software And Standards

validate user requirements while delivering capability to the user incrementally and reducing the document and review process overhead. The rapid prototyping methodology is somewhat similar to the early stages of the Spiral Model methodology. In fact, rapid prototyping is envisioned to play a significant role in the evolutionary development methodologies. Unfortunately, rapid prototyping has resulted in the development of very useful "throw away code" because of a lack of robustness, maintainability and expandability. Overcoming this deficiency in rapid prototyping is the primary basis of the RDM. Therefore, the objectives of the RDM are to preserve the flexibility and responsiveness of rapid prototyping and to deliver software meeting the requirements of good design and accepted coding standards, resulting in a maintainable system.

While RDM is not rapid prototyping with some added documentation; it is also not repetitive rapid executions of the Waterfall methodology, as in the Incremental Waterfall methodology. It is the iterative nature of the RDM with increasing documentation and formal reviews with each delivery that permits a departure from the Waterfall methodology and yet results in the end with a product identical to that developed using the Waterfall approach, from a maintainability perspective.

The RDM is a software development methodology and project management approach with a specific set of tenets. The basic tenets of the RDM that have been recognized to date are the following:

1. The system's functionality can be delivered and used incrementally. Each delivery must consist of a full system and not just pieces of the final system. Subsequent deliveries enhance or add functionality to previous deliveries.
2. The operational use of each incremental delivery provides active feedback of requirements into future incremental deliveries.
3. Users of the system will interact with developers extensively during the development of each incremental delivery. Users are active participants in all phases of the system development.
4. Users must agree that deficiencies in one delivery can be fixed in future deliveries.
5. The development cycle and the documents become progressively more formal as incremental deliveries are made.
6. The system is developed from project inception with an overall architecture that is sufficiently modular to allow for an evolving system development.

As currently budgeted, the EUCOM Project will proceed with five incremental deliveries resulting in the Early Operational Capability (EOC) at the end of FY93. The current schedule of deliveries is shown in Figure 4. One incremental delivery overlaps with the preceding and the succeeding deliveries in order to balance the staff and to achieve a 9 to 12 month interval between deliveries. Maintaining a nearly level staffing profile is necessary in order to achieve multiple incremental deliveries. Additionally, through successive deliveries, the users gain increasing understanding of computer automation benefits, which allow them to better articulate requirements and the developers gain more experience and expertise with the development environment. It is the refinement of requirements and increasing developer expertise through successive deliveries that permit greater detail of documentation and greater formality.

Figure 5 depicts the RDM process. The Project starts with an initial phase to establish an overall framework. This includes establishment of the overall requirements, schedule, budget and conceptual architecture/design. Each delivery starts with a planning and requirements phase. An agreement is reached with the Sponsor and the user on the functional requirements, schedule and budget for the delivery. Segment (sub-system) level partitioning, requirements definition and top level designs are then prepared. Also, detailed bottoms up costing and scheduling are performed. This results in a formal commitment review within JPL and a detailed commitment to the Sponsor/User. Implementation then proceeds similar to that of the Waterfall implementation phases, broken down into design, code and unit test and system test sub-phases. However, testing of the incremental system is somewhat different from the Waterfall approach, in that operational usability of the system is emphasized. This contrasts with an extensive acceptance test period emphasizing satisfaction of system requirements. Prior to installation and integration at the users' site, a formal Delivery Pre-Integration Review is held within JPL.

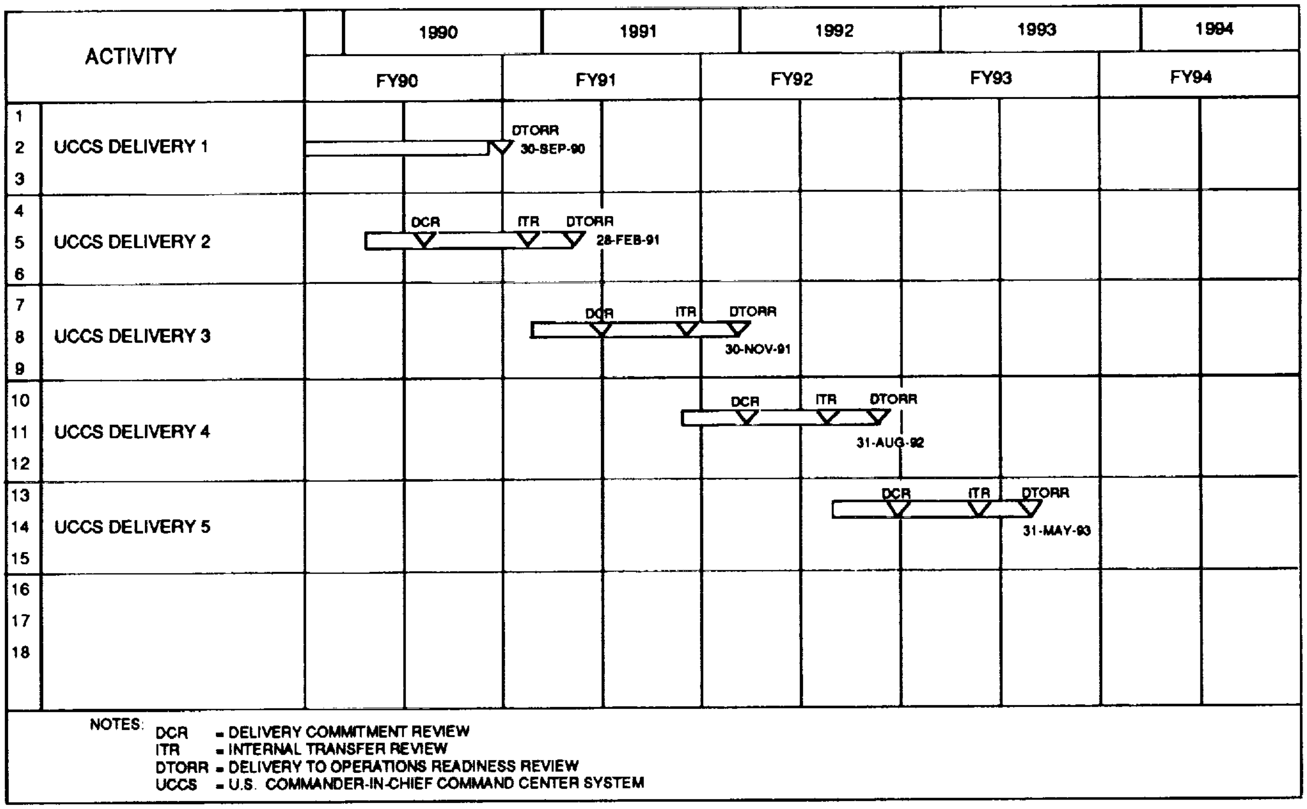


Figure 4. EUCOM Command Center Project Schedule

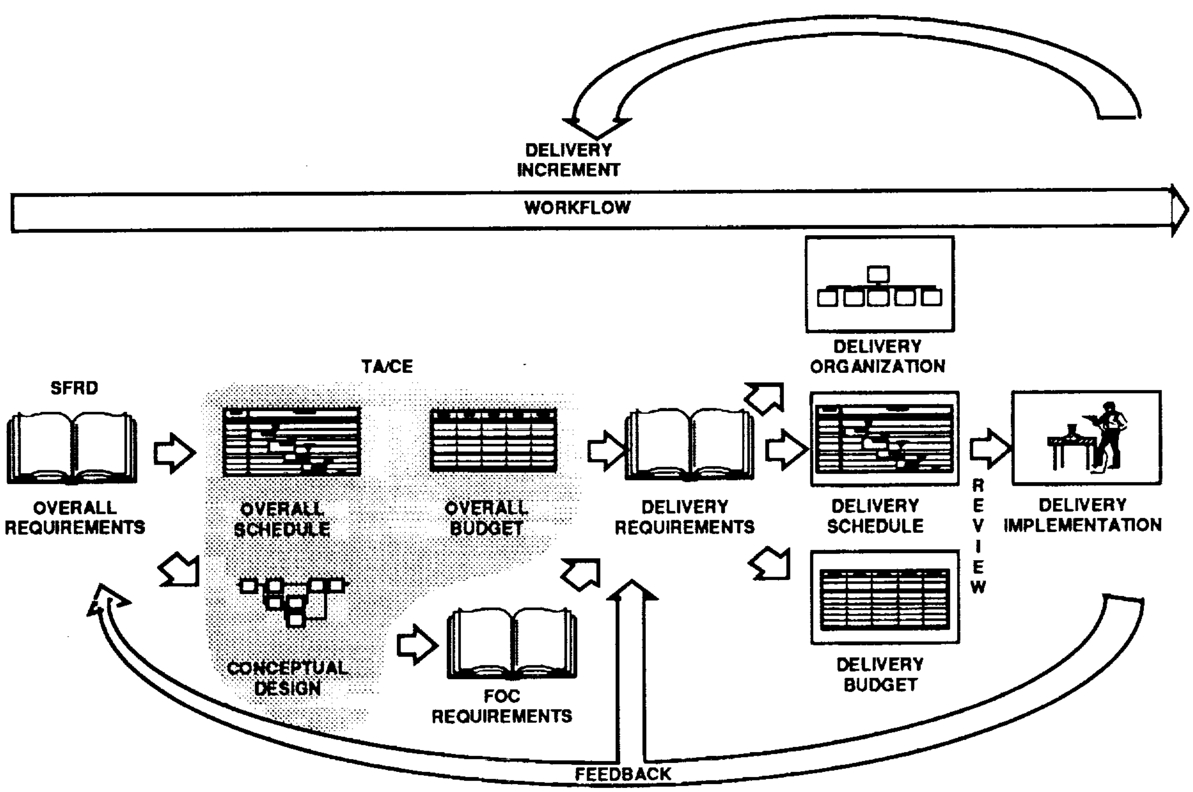


Figure 5. Rapid Development Process

As previously indicated, the users are involved extensively in the development of the system. The users state the requirements, through feedback of previous deliveries and discussions about future deliveries, at the beginning of each delivery. The users review all documents including System and Software Requirements Specifications and Software Design Documents. The users are part of the planning process setting priorities of requirements for each delivery cycle. The users review the implementation while it is in progress by frequent interaction with the developers (i.e. review of user interface storyboards and other design material, user interface prototypes, and partially working code). The users participate and make the final decision regarding cost/capability tradeoffs. The users are the final testers of the system as system operators/abusers. The users, through an established change request process, change the requirements for future incremental deliveries based on their operational experience. Finally, the users accept (or reject) the system. Since user participation is an integral part of the RDM, the user assumes "ownership" of the system early in the process.

The increased user participation is beneficial only if the users understand that deficiencies in one delivery will be addressed either as maintenance updates to the current delivery or included in future deliveries. On-site sustaining engineering support and timely and reliable developer-to-user feedback are provided to maintain a good working relationship with the user community.

THE Ada PROGRAMMING LANGUAGE

The Ada programming language was developed in response to a perception on the part of the United States Department of Defense of a software *crisis* within the department's large scale complex computer software developments [6]. These complex software developments were perceived to lack maintainability, transportability to other hardware platforms, and lack support for code reuse. They were also difficult to integrate and manage due to a proliferation of source languages.

Ada was developed to address problems associated with large scale computer system developments and ultimately reducing total program lifecycle costs. In 1984, the DoD mandated that all future mission-critical software developments be developed in the Ada programming language [6]. The objectives of the Ada programming language were to: 1) develop a single, common programming language for all large scale software developments, complete with software development tools, to aid in all aspects of Ada software development, thus increasing the experience base of software programmers, 2) embody and enforce modern software engineering goals, such as program modifiability, efficiency, reliability and understandability, and 3) develop one language which would facilitate the transfer of software to different hardware platforms and operating systems by isolating the hardware dependent features of the language [6].

The Ada programming language was designed to embody and enforce modern software engineering principles such as data abstraction, information hiding, program modularity, localization, uniformity, completeness and comfirmability [6].

THE ECCP SOFTWARE DEVELOPMENT LIFECYCLE

The EUCOM Project planners chose Ada as the standard programming language for all Project software development primarily due to the DoD Ada mandate and for the hope of realizing the benefits described briefly in the above paragraphs. However, the use of Ada required some modifications to the implementation phase definition of the RDM in the areas of sub-phase scheduling and costing, development tools implementation, and Ada coding standards.

Implementation Phase Scheduling and Costing

The EUCOM Project uses a design-code-test ratio of 45-25-30. That is 45 percent of the Project's implementation phase is spent on preliminary and detailed design activities, 25 percent of the development schedule is spent on code and unit test, and the remaining 30 percent of schedule is spent on integration and requirements testing of the software (not including user tests as required for the RDM).

A higher percentage of implementation schedule is spent during the design phase than the normal 19 percent suggested by B. Boehm [ref. 1 pg. 65 for 32 KDSI] due to additional effort required to define all data types and develop package specifications. (In Ada, a collection of like procedures and functions is encapsulated into the Ada "Package". Each package consists of the package specification, which is the visible part of the package, and the body, which is the actual code.) Ada is a strongly typed language, and thus requires more time to define and negotiate all required data structures and type definitions.

Less time is required for the ECCP coding phase than the normal 55 percent as suggested by B. Boehm, because most sub-program interfaces are negotiated and agreed to during the design phase.

The test phase is different than traditional developments in that very little test time is used to find data type errors and sub-program interface mismatches, found during the design and coding phases. Also, recall one tenet of the RDM is that the user participates in all phases of the development, including the test phase.

In Ada, programming elements developed by one individual programmer, but used by others, such as procedure argument lists, data types, and external package references, are fixed in the package specification. For a large scale software development, this rigid definition of software program elements speeds usage of those elements by other members of the development team and speeds testing of sub-program interfaces.

Software Development Tools

As part of the ECCP's planning, a Rational Ada Development System was purchased from the Rational Corporation. The Rational is a complete Ada development environment with a 2167A document generator, an extensive configuration management system, a sophisticated language sensitive editor and Ada compiler. Code is first generated and compiled on the Rational. When the developer has achieved a successful compilation, the source code is transferred to the target environment (the DEC Station 3100 with Ultrix version 4.1) for unit testing. This transfer is managed automatically by the Rational.

The Rational development system, being dedicated to development of Ada software, has many features which support Ada programming that save a great deal of development time. One of the routine operations that a programmer performs during the coding phase is to refer to data type definitions. Although this may sound simple enough, with a strongly typed language, this is often time consuming and fatiguing. Most data type definitions are actually compound data types, with each component of the compound type itself being of a specific type. While coding, it is frequently necessary to examine several data type definitions at once, and these are commonly defined in different packages. The Rational system provides many shortcuts for viewing these definitions, saving a substantial amount of programmer time.

The Rational keeps track of highly detailed information regarding the interdependence of procedures, type definitions, and packages. This feature can be used to quickly identify, in a very specific way, any program entities which depend on another entity. The programmer uses this capability to determine the impact of any proposed changes, and to quickly traverse between the parts of the affected modules, speeding the process of making the software consistent with the modified portion.

Some other features of the Rational include syntactic assistance, which provides on-line Ada statement syntax completion and/or validation, semantic assistance which provides procedure argument completion and/or validation, incremental compilation which greatly reduces compile times of complex software programs, and an integrated configuration management system which provides automatic version and configuration management as part of the code generating process.

Since the RDM is based on multiple deliveries to the sponsor, feedback from the user is provided during the requirements definition and implementation phases of the next delivery. This feedback comes in the form of problem reports and requests for functionality enhancements. Because this feedback must be addressed during the development of the next delivery, configuration management plays an extremely important role in all phases of each delivery, regardless of the size of the programming staff, or number of lines of code to be managed. It is this overlapping of delivery developments and continual user requests for software fixes and enhancements that necessitate a great need for configuration management. For example, between the ECCP's last and current deliveries, the main database structure has changed to one that more adequately reflects real-world situations. But the previous delivery must still be supported at the same time.

The ECCP software configuration management process is well-defined and strictly enforced. Without this strong configuration management presence, much time would be lost to version mismatches and the resulting recompilation times. No time has been lost to date due to a software build with an incorrect package version.

Project Ada Coding Standards

The EUCOM Project has developed a set of coding standards which are tailored to the lifecycle discussed in this paper and take advantage of the benefits of the Ada programming language. Some examples of the Project's Ada coding standards are strong enforcement of Ada typing restrictions, utilization of Ada's run-time checking of type ranges, development of an error handling policy which takes advantage of Ada standard exception handling for calling applications but allows simple returned statuses from the system support layers to the calling application, and use of Ada generics and common libraries to facilitate code reuse and program modularity throughout the Project.

The ECCP Ada coding standards contain naming conventions to ensure that package and procedure names are meaningful and restrict the usage of the "use" clause, requiring the explicit use of reference labels. Both of these standards are aimed at providing more readable software.

In addition, all user interface routines are contained in standard libraries, and all applications are required to use these libraries. This enforces the Project's Man-Machine Interface policy at the same time.

The ECCP's coding standards facilitate good software engineering practices, which is required by the RDM for code expandability and modifiability from one delivery to the next.

BENEFITS AND CHALLENGES OF ECCP SOFTWARE DEVELOPMENT LIFECYCLE

Benefits

The benefits of RDM itself are many. Budgets and requirements for each increment can be fixed, thereby insulating the increment from the effects of overall funding and requirements volatility. The methodology is responsive to funding gyrations and requirements changes as a result of the evolution of the users' understanding of the system and the associated procedures for using it. The users gain a satisfaction with the system because of their "ownership" through participation in the development. Further, the RDM incrementally introduces changes in the operations organization rather than revolutionizing it with the introduction of a complete new system. Finally, RDM delivers capability faster than the turnover of user personnel - a major benefit for military systems.

One goal of the implementation phase of the RDM is to provide any reduction in time expenditure possible throughout each of the sub-phases. The ECCP has been able to provide these time savings through the use of the Ada programming language and its support tools. Time savings have been realized in the areas of software configuration management, using Ada to capture the design early in the design phase, making use of the language's extensive compile and run-time checks and using the language's inherent support for program modularity. Each of these areas is discussed below.

For the ECCP, each delivery increment was anticipated to be relatively small. Less than seventy thousand lines of code were developed for each increment by 12 to 14 actual software developers. The management of software change over multiple deliveries and an ever-changing and increasing line of code count increase the complexity of the software development environment. By using the Ada programming language and enforcing rigid configuration management standards, the difficult process of managing change to past and present software deliveries is minimized, almost to the point of being void of problems. In the accelerated schedule demanded by the RDM and with minimal resources available for configuration management functions, the combination of a strong configuration management policy, use of automated tools such as the Rational and the modularity features of the language all combine to save substantial programmer time, when compared with less automated or manual tools and the use of other languages lacking built-in modularity support.

The ECCP has designed its development process to exploit the advantages of the Ada language during each phase of the development. For the design phase, emphasis was placed on developing Ada package specifications, containing actual Ada code, rather than pseudo-code. By moving directly to compilable Ada statements, the step of translating pseudo-code into High Level language (HOL) statements was avoided. We believe that the inherent readability of Ada makes skipping the pseudo-code step possible.

The benefits of using Ada in place of pseudo-code are greater than just saving the time it takes to translate pseudo-code to HOL statements. By using Ada, the interdependencies of the compilation units are rigorously defined. Therefore, it is possible in the early design phase to determine whether there are any undesirable dependencies among the compilation units, and to resolve them in this early phase of development. By clearly and rigorously defining the architecture with actual Ada package specifications, architectural problems with the design can be identified and corrected before full implementation begins. This reduces the likelihood that significant portions of code will need to be rewritten without incurring the additional time for pseudo-code development.

During the actual implementation phase, Ada provides extensive error checking and error messages when compiling. Since procedures and functions have rigorously defined argument lists, and the arguments themselves are strongly typed, Ada can identify many common coding mistakes, and notify the programmer at the time of compilation. This provides some cost savings of time over the more traditional cycle of compiling, building, testing, and debugging without these compile and run-time checks, built into the language. This savings is especially important for software developed on a foreign host, such as the Rational system, and then downloaded to a target system. For such systems, the code generally needs to be compiled on both the development and the target system, in addition to the extra step of moving the source files from the host to the target. The sooner the programmer is aware of a problem, the more time is saved.

Another benefit of this approach is that the EUCOM Project is able to field versions of the system every nine to twelve months with each delivery providing significant new and modified capabilities to the operational environment. With the first decision support delivery in March of 1991, many initial requirements have undergone revision and redefinition based on the user's experience with the operational system. The Project's strong bias toward program modularity and its extensive use of common libraries permit significant code reuse from delivery to delivery. While some applications have been extensively reworked, the common libraries have been only enhanced and not re-designed. This has allowed us to provide significant new functionality in the next delivery, now scheduled for March, 1992 and at the same time making extensive modifications to some existing applications.

Challenges

While early, incremental operational capabilities are well received by the user communities, our experience has shown that government review agencies are reluctant to accept the methodology and its initially less rigorous testing and review processes. Our challenge has been to allow close participation with government Quality Assurance organizations, which tend to require complete documentation sets and lengthy review cycles, while still maintaining the accelerated delivery schedule dictated by the methodology.

One challenging aspect in the selection of Ada has been that Ada requires total control of the program execution environment. By controlling the program execution environment, all the run-time benefits designed into the programming language can be realized. Unfortunately, this philosophy conflicts with other operating environments which tend to control the environment. Examples of conflicting controlling environments are the UNIX operating system and the X Window Graphical User Interface. For example, we have demonstrated that the use of multiple Ada tasks in a single UNIX process have yielded unpredictable and erroneous results in our target environment. Careful management of the interaction between each of these environments is required for a successful implementation of these government standards. Our solution to the problem has been to restrict the use of Ada tasking by making each Ada task a separate Unix process.

One pitfall associated with this approach is the increased cost of supporting the operational system due to its early fielding. The operational support of the system has been addressed through a combination of on-site and military support teams. USEUCOM has provided a 24 hour operations support team during the first delivery in addition to the on-site JPL support team. The JPL on-site support team provides assistance for technically challenging problems, not day-to-day operational problems. Additionally, USEUCOM is developing an on-site user support team, responsible for day-to-day user training, and general user support. The Command and JPL have accepted this cost as necessary for the continued support of the user community.

Another pitfall associated with this approach is the possibility of architectural changes due to future user requirements which were not anticipated during the original architectural design. This has been addressed by the modularity of the system architecture, minimizing and isolating the impact of future architectural changes. Prior to each commitment review, architectural changes are factored into the committed work plans and agreed upon by the sponsor/user.

RESULTS

The ECCP is midway through development of the key elements of a modern command and control system, adhering to government-mandated (and sometimes conflicting) software standards. The Project's alternative methodology has the advantage of delivering substantial capabilities to the user much earlier than other methodologies without sacrificing lifecycle development quality or rigor. During each delivery, the duration allocated for the code and test phases have been reduced allowing more capability to be developed and delivered.

The Project has developed portable software modules which can be applied to any distributed computing environment requiring monitor and control, client-server database driven decision support applications in a command and control atmosphere. The use of standards, development of reusable software "modules" and the early delivery of an operational system has proven to be very beneficial in the current limited budget environment.

ACKNOWLEDGEMENT

The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

REFERENCES

1. Boehm, B., Software Engineering Economics, (Englewood Cliffs, NJ: Prentice-Hall, 1981).

2. Firth, R., Wood, B., Pethia R., Roberts, L., Mosley, V., Doice, T., "A Classification Scheme for Software Development Methods," Technical Report, Software Engineering Institute, Carnegie-Mellon University (November 1987).

3. Scacci, W., "Models of Software Evolution: Life Cycle and Process," SEI Curriculum Module SEI-CM-10-1,0 (October 1987).

4. Sodhi, J., "Managing Ada Projects," (Tab Books, 1990).

5. "JPL Software Management Standards Package" the Jet Propulsion Laboratory, D-4000 (Internal Document) Version 3.0 (December 1988).

6. Grady Booch, "Software Engineering with Ada," (Benjamin/Cummings, 1986).

7. Military Standard, "Defense System Software Development, DoD-STD-2167A," (Washington D.C., June 1988).

N92-22470

ADVANCES IN KNOWLEDGE-BASED SOFTWARE ENGINEERING

Walt Truszkowski
Head, Automation Technology Section
Code 522.3
Goddard Space Flight Center
Greenbelt, MD 20771

ABSTRACT

The underlying hypothesis of the work reported on in this paper is that a rigorous and compre- hensive software reuse methodology can bring about a more effective and efficient utilization of constrained resources in the development of large-scale software systems by both the Government and industry. It is also believed that correct use of this type of software engineering methodology can significantly contribute to the higher levels of reliability that will be required of future operational systems.

This paper presents an overview and discussion of current research in the development and application of two systems that support a rigorous reuse paradigm: the Knowledge-Based Software Engineering Environment (KBSEE) and the Knowledge Acquisition for the Preservation of Tradeoffs and Underlying Rationales (KAPTUR) systems. The paper concentrates on a presentation of operational scenarios which highlight the major functional capabilities of the two systems.

THE KNOWLEDGE-BASED SOFTWARE ENGINEERING ENVIRONMENT (KBSEE)

The KBSEE (refs.1, 2) is one of the systems that is being used to substantiate the hypothesis stated above. This system currently supports a comprehensive software specification reuse capability. A central concept for the KBSEE is the concept of a domain. A domain, in our context, is any class of related objects. These objects can be as diverse as cruise control systems for cars, elevator systems, or spacecraft command and control systems (in fact all three of these domains have been used to demonstrate the functionality of the KBSEE system to date). A high-level view of an end-to-end (reusable object definition to target system specification) KBSEE scenario is as follows. A domain model is developed as part of an initial exercise in populating the reuse knowledge base and, in an internal standard form, is stored in the KBSEE's reuse repository. When the requirement for a new instance, or target system/application, of that model or portion thereof arises, the KBSEE provides the software engineer with the capabilities to tailor the existing model to meet the target application's requirements and constraints. The output from the current system is the specification for the required target system. If, in the development of the specifications for the new target system, some of the requirements or constraints cannot be met by the current domain model a feedback to the domain modeling capabilities of the KBSEE allows the current model to be modified in order that the new requirements or constraints may be reflected. This process thus allows the domain model to evolve to a richer model. This phenomenon gives rise to the concept of an evolutionary domain life cycle. The Figure 1. graphically illustrates the major processes supporting the domain modeling and target system generation stages of the KBSEE. Model-wise the KBSEE is highly object-oriented and there is a similarity in concept between the tailoring of a domain model and the use of class definitions in arriving at instances of objects.

A major experiment has been conducted to better explore the applicability of the KBSEE in support of the development of specifications for spacecraft control center software components. Based on an extensive domain analysis of several control center software systems a generic control center software system model was established. This model was used as the basis for the definition of reusable software specifications for target applications in the control center environment. As depicted in the following two figures the KBSEE provides two basic types of capabilities: creation of a domain specification and generation of target system specifications.

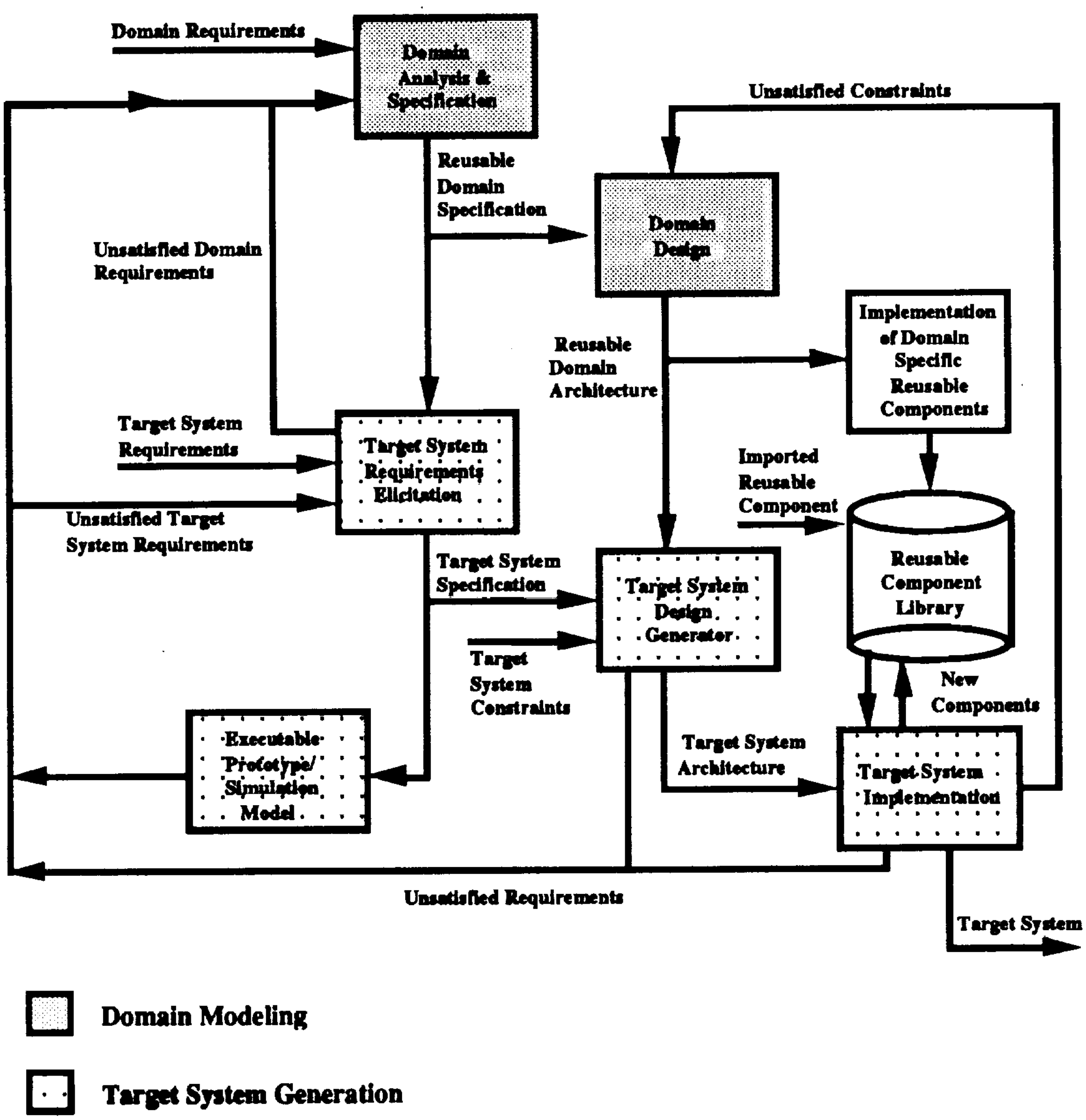


Figure 1. The Evolutionary Domain Life Cycle Model

The Figure 2. illustrates the object-oriented domain model specification and reusable component generation capabilities provided by the KBSEE. In the development of reusable objects in a domain of interest the software engineer is provided capability by the KBSEE to develop domain component specifications in a number of representations, namely:

1. object communication diagrams which are hierarchically structured and show how objects communicate with each other through the mechanism of message passing,
2. aggregation hierarchies which supports the decomposition of complex aggregate objects into less complex objects eventually leading to simple objects at the leaves of the hierarchy,
3. generalization/specialization hierarchies which support the IS-A relationship and inheritance relationships between classes and instance objects, and
4. state transition diagrams which reflect the fact that objects are sequential processes which may be represented by finite state machines and documented by state transitions.

Once the specifications have been developed by the software engineer the multiple representations are checked for consistency among themselves. If there are any inconsistencies these are brought to the attention of the software engineer for corrective action. Once all the inputs are consistent they are transformed into an internal format for storage in and later retrieval from the domain object repository.

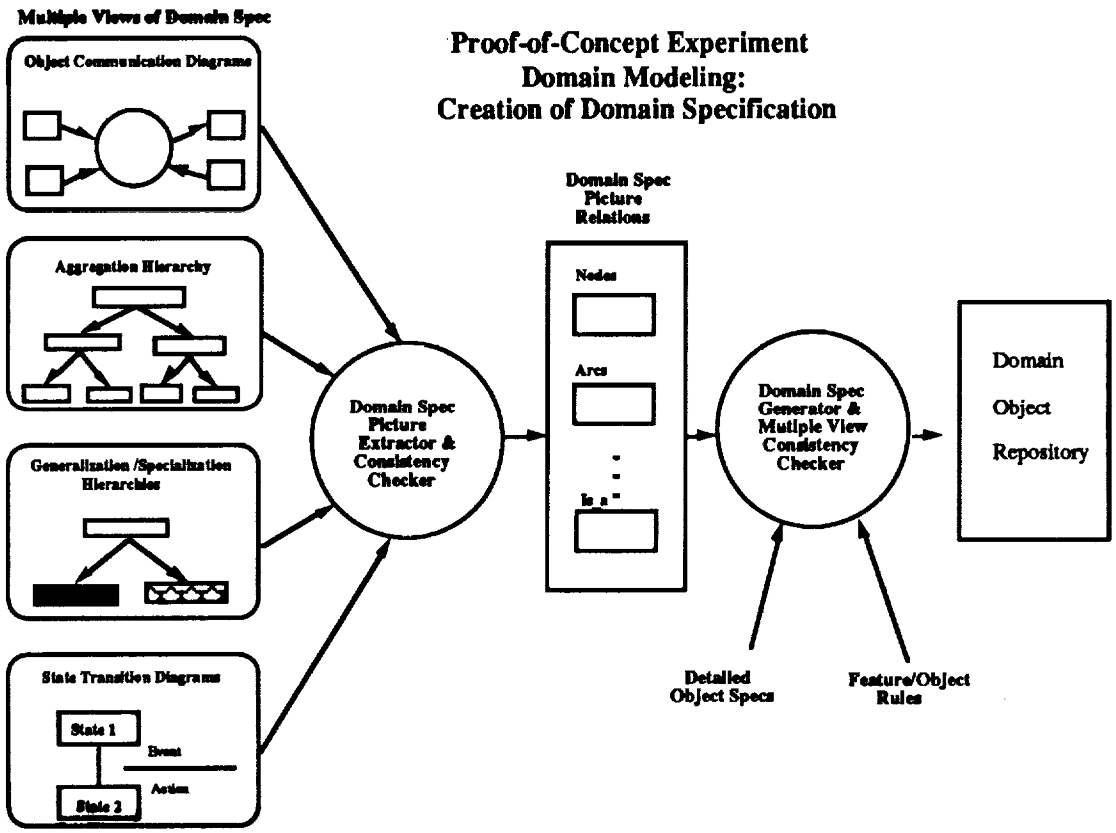


Figure 2. Creation of a Domain Specification and Reusable Components

The following Figure 3. depicts the major steps in the process of generating target system specifications using the KBSEE. In many ways it is the reverse process of the domain modeling processes depicted in Figure 2. A typical scenario for this aspect of the KBSEE would be as follows. A software engineer would invoke the KBSEE with a specific target system in mind. The KBSEE's knowledge elicitation component, KBRET, would engage the software engineer in an interactive exchange during which time the requirements and constraints for the target system would be made known to the KBSEE. This process currently involves the engineer selecting from requirements and constraints currently supported by the appropriate domain model in the KBSEE knowledge base of reusable objects. If the target system requires a feature not currently satisfied by the domain model the KBSEE, through a feedback mechanism, allows the feature to be added to the domain model through the processes depicted in Figure 2. Once all of the target system objects have been defined the KBSEE, through its target system picture generator, creates the multiple views of the target system specification which is the current output of the system.

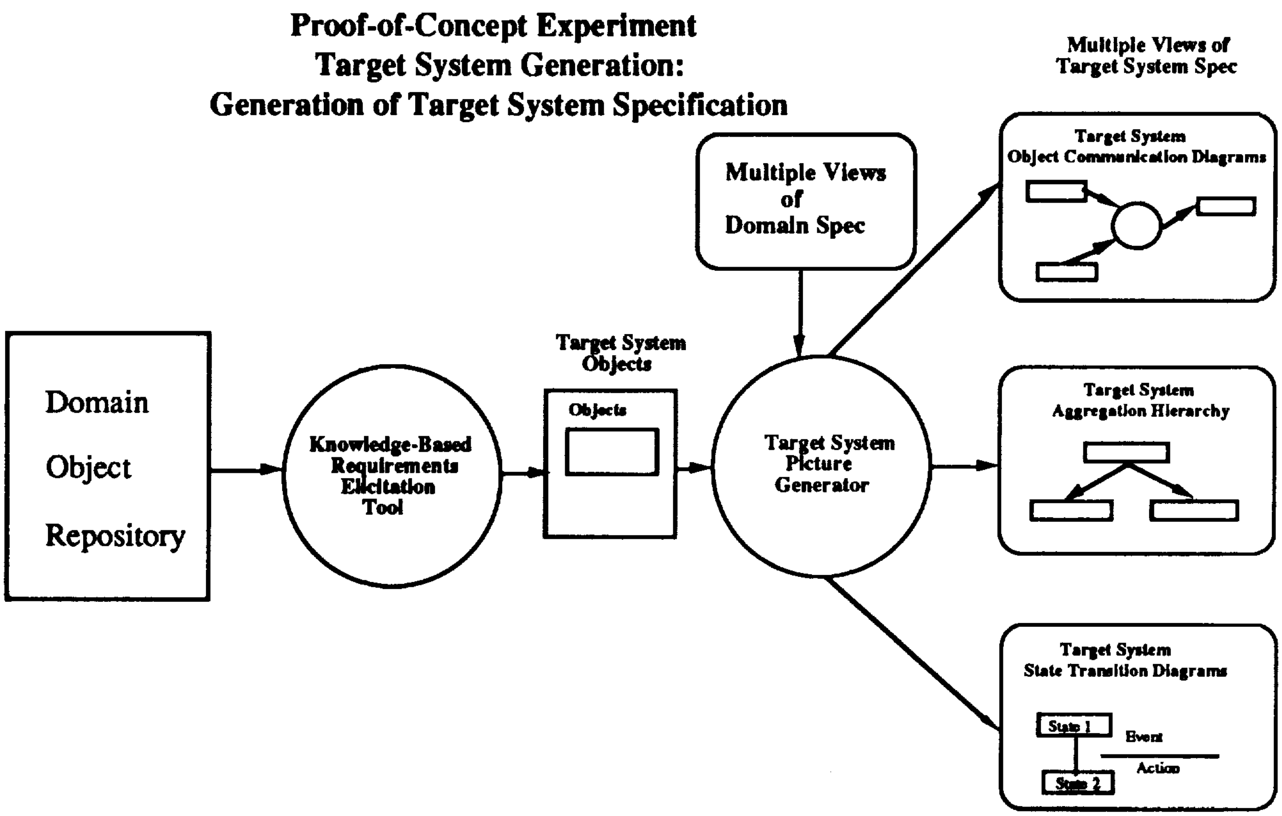


Figure 3. Use of the KBSEE in Target System Specification Generation

As mentioned above, a major innovative feature in the KBSEE's support for target system specification is depicted in the above figure in the left-most circle. It is the Knowledge-Based Requirements Elicitation Tool (KBRET) and a fuller discussion of this tool follows. This tool supports the development of target system specifications by entering into an interactive session with the target system designer. During the session KBRET elicits target system requirements by having the designer select desired features and object types associated with the domain model. KBRET queries the object repository to obtain the knowledge required for its reasoning in support of the target system specification generation process. Modern browsing techniques are being used by KBRET to support this access.The Figure 4. illustrates the major components of the KBRET tool. A major attempt has been made to ensure that the user's interface to the KBSEE through the KBRET tool is as user accommodating as possible. An interesting and powerful aspect of the KBRET architecture is the organization of its knowledge base. The KBRET knowledge base is divided into domain dependent and domain independent portions. The domain independent portion support such activities as browsing, the selection and/or deletion of target system features by the user, a check for dependencies among selected features based on the domain model, and finally an invocation of the target system generator process. All of these activities apply equally to any domain selected. The domain dependent portion makes extensive use of the highly structures object-oriented reuse knowledge base and provides the data and information required to support the domain-specific specification development process.

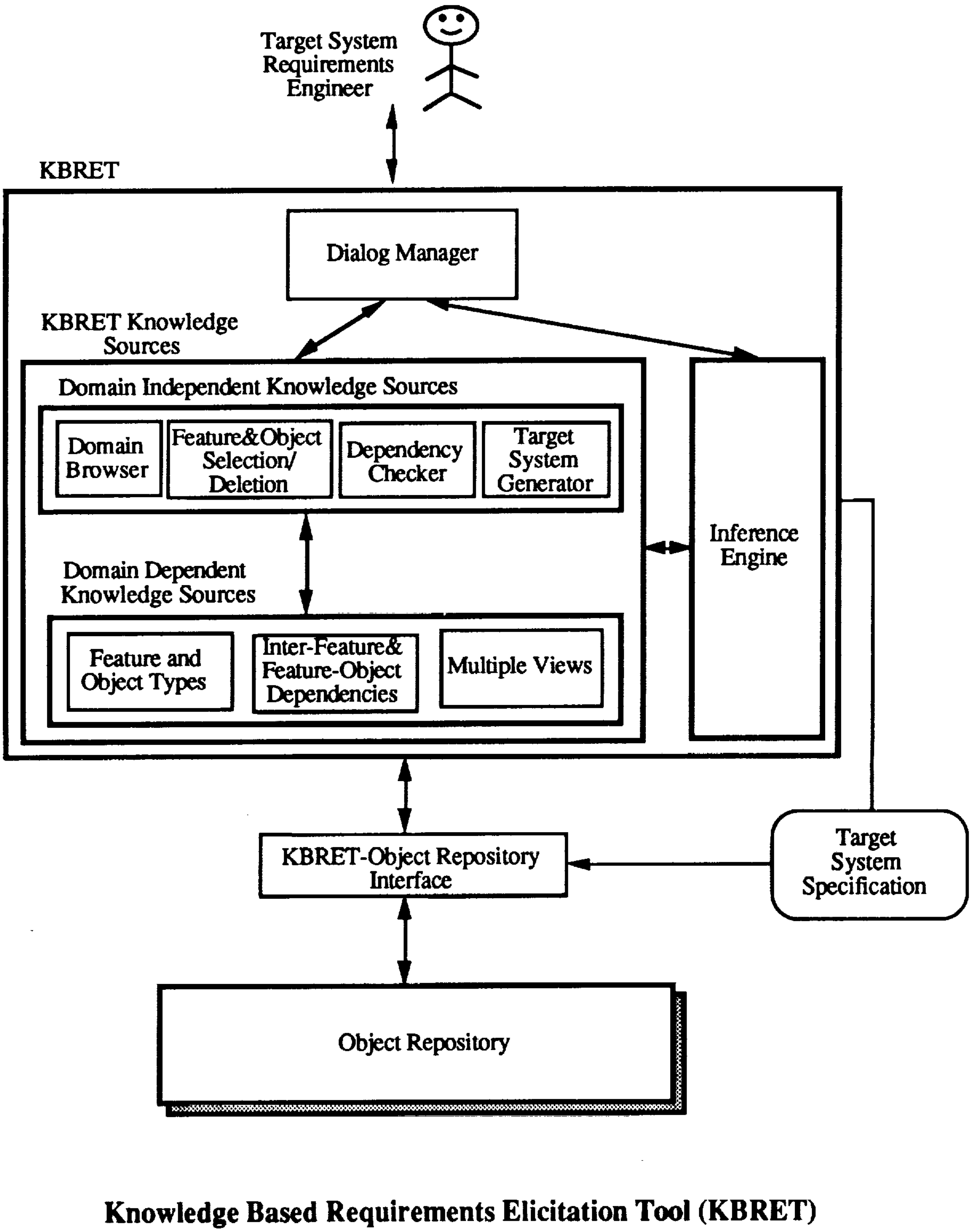


Figure 4. Component View of the KBRET Tool

The present version of the KBSEE is implemented on a Sun and utilizes Software through Pictures as the multiple viewpoint graphical editor, Eiffel as the basis for the object repository, CLIPS for the KBRET knowledge elicitation component, and TAE+ for the KBSEE graphical interfaces.

KNOWLEDGE ACQUISITION FOR THE PRESERVATION OF TRADEOFFS AND UNDERLYING RATIONALES (KAPTUR)

The KAPTUR system (ref. 3) shares many of the goals as the KBSEE. The KAPTUR system is intended to support systematic reuse of knowledge and artifacts throughout the software development life-cycle. The main contribution of KAPTUR is the support it provides for the evaluation of potentially reusable artifacts, enabling the developer to make intelligent choices among the possibilities. KAPTUR is intended to provide as much information as possible, in an easily accessible form, to help clarify whether a given artifact is suitable for reuse in a given context.

KAPTUR is intended to preserve knowledge that is required or generated during the development process, but that is often lost because it is contextual, i.e., it does not appear directly in the end-products of the development process. Such knowledge includes issues that were raised during development, alternatives that were considered, and the reasons that were used to choose one alternative over another. Contextual information , in our sense, is usually only maintained as a memory in a developer's mind. As time passes, the memories become more vague and individuals become unavailable, and eventually the knowledge is lost. KAPTUR seeks to mitigate this process of information attrition by recording and organizing contextual knowledge as it is generated. From the lessons learned from previous development efforts, current developers can improve their insight into the problems at hand and their possible solutions.

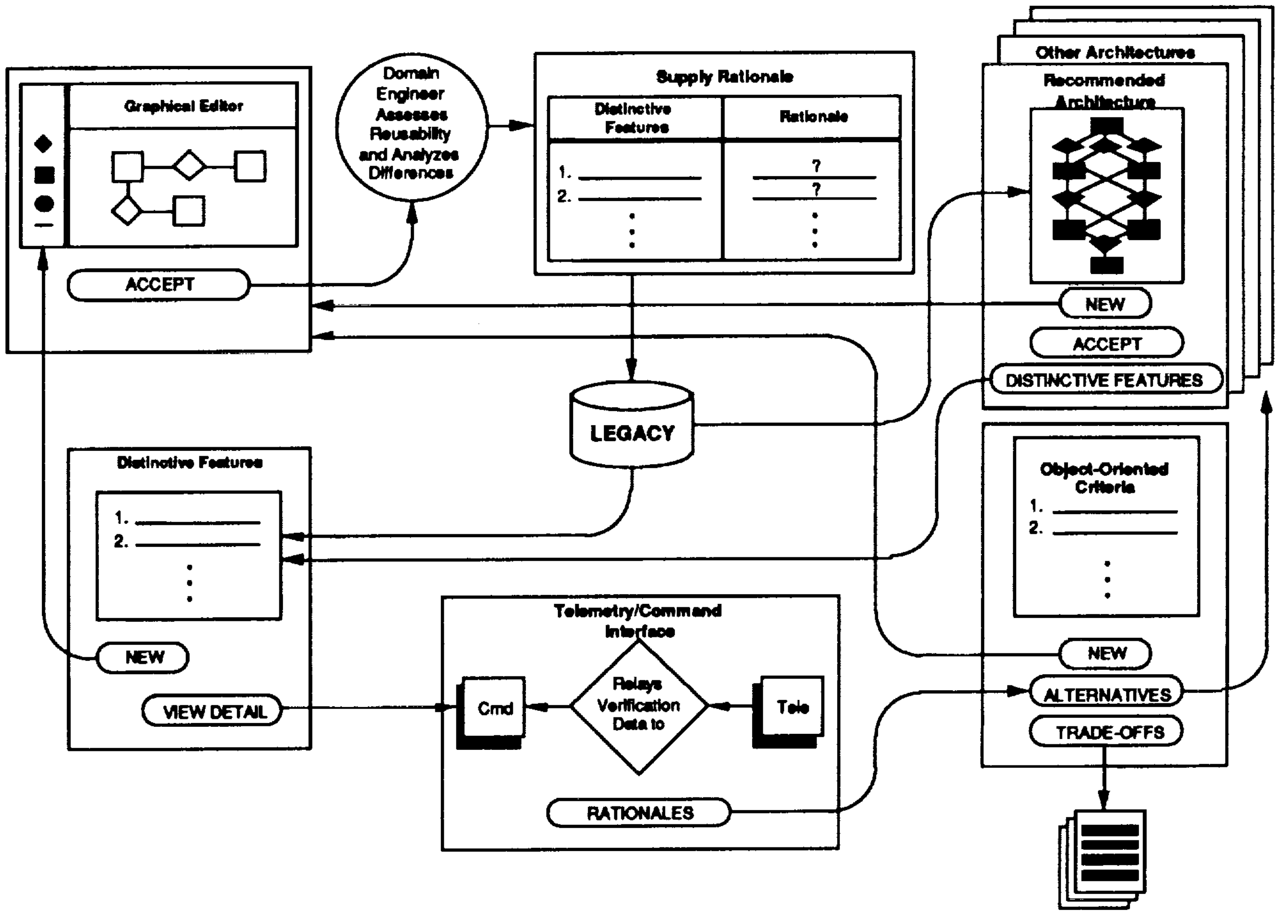
The following Figure 5. illustrates the underlying operational philosophy of KAPTUR. Specifically it shows how KAPTUR would be used to explore alternative software architectures for a control center application processor. In the center of the diagram there is a knowledge base, called the Legacy, which contains information about the application domain. In our case the application domain is control center software systems. The information includes recommended architectures and information about previously developed systems. In the scenario depicted in Figure 5. the developer has available a set of software requirements and wants to begin defining an applications process to meet these requirements. The developer sits down at the KAPTUR workstation and issues a command whose meaning is something like the following: "I want to develop a control center applications process. Show me what they look like." In response the KAPTUR system displays the recommended generic architecture (upper right-hand box) as well as a stack of alternative architectures related to this requirement. Upon examining the recommended architecture the developer has the following options:

- examine the distinctive features of the recommended architecture,
- examine the alternatives by clicking on one of the windows behind the recommended architecture,
- define a new architecture,
- accept the recommended architecture.

The distinctive features of an architecture are those that are different from common practice or the recommended approach, or that represent a non-trivial decision about a significant issue. It is the prime purpose of KAPTUR to preserve the knowledge and analysis of the decisions associated with the distinctive features. Distinctive features may correspond to specific portions of an architecture (e.g., the interface between two subsystems) or they may represent some aspect of the architecture as a whole (e.g., the distribution of initialization functions to all subsystems of a system).

If the developer selects Distinctive Features, KAPTUR will list the distinctive features of the architecture being displayed., and will allow the developer to select one or more of these features. KAPTUR will then display a representation of the distinctive feature(s). In effect, the developer is afforded the opportunity to zoom into a view of a particular feature of an architecture. This is illustrated in the bottom-middle box in Figure 5.

The developer can then examine the Rationales for this feature, i.e., the reasoning underlying the decision that the feature represents. In the lower right-hand box in the Figure 5. the rationales are represented as a list of object-oriented design criteria that might underlie the decision. From this screen the developer can request even more detailed explanations by asking to view Trade-Offs that were considered in making the decision. The developer can also ask to view Alternatives to this decision, i.e., other systems that do not possess this feature because a different decision was made.



KAPTUR may be used to explore alternative software architectures

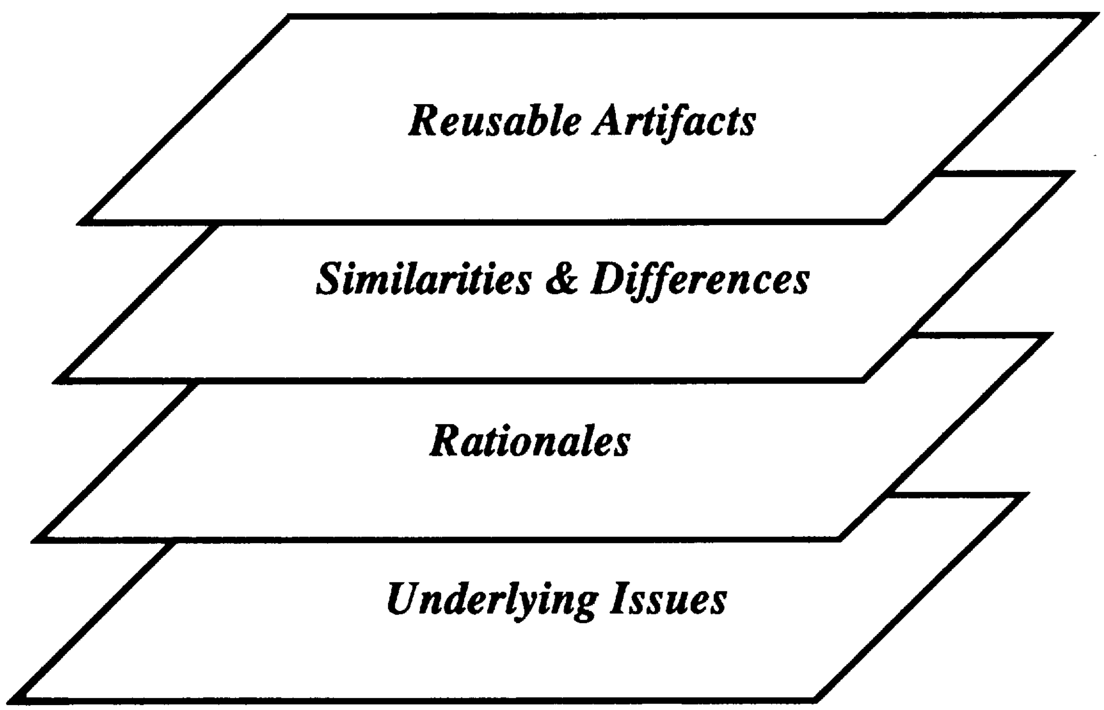
Figure 5. A Scenario View of the KAPTUR System

If the developer selects New (from either the Recommended Architecture or Alternative Architecture screens) a graphical editor will be invoked to allow the interactive definition of a new architecture. The definition of a new architecture need not start from scratch. A clipboard capability in KAPTUR allows the developer to select portions of the recommended and/or alternative architectures for inclusion into the new architecture.

Once a new architecture has been defined the developer will perform an analysis to determine the distinctive features of the new architecture, i.e., the ways in which it significantly differs from the recommended architecture. For each identified distinctive feature KAPTUR will prompt the developer to enter one or more rationales justifying the feature. This is shown in the top-middle box in Figure 5.

The new architecture, together with its rationales, then become part of the Legacy of the domain and will appear in the Alternatives list when KAPTUR is next used. This is how the evolution of domain requirements and solutions is captured in the knowledge base.

The knowledge in KAPTUR is stratified into four layers as depicted in Figure 6. It is through this multi-leveled knowledge base that KAPTUR is able to support the mechanisms for design knowledge capture within a robust reuse environment.



Layers in KAPTUR's Knowledge Base

Figure 6. Structure of the Underlying Knowledge Base for KAPTUR

SUMMARY AND CONCLUSIONS

The two systems described in this paper have been developed over a period of three years to serve as testbeds for the prototyping and evaluation of knowledge-based and advanced software engineering concepts needed to support a rigorous software reuse paradigm. Among the major concepts studied have been those associated with:

- representation of reusable software specifications
- consistency checking among various specification formalisms
- knowledge-based approaches for interactive requirements elicitation
- mechanisms for design knowledge capture
- hierarchical structuring of design knowledge
- knowledge-based browsing techniques
- user/system interaction
- object-oriented knowledge bases

Both of these systems are currently being used to focus on the issues associated with software reuse in the context of spacecraft control center software system specifications. As NASA missions become more complex, long-lived, and increasingly expensive the developmental and cost-savings benefits that can be derived from a well formulated reuse methodology take on added significance. Especially for those programs that have a long projected lifetime, the need for establishing and maintaining a "corporate memory" of reusable components and system development rationales becomes critical for an effective sustaining engineering activity. Over the next year both of these systems will be field-tested on real-time control center software development projects to help in further evaluating their effectiveness in operational settings.

We feel strongly that the concepts embodied in systems like the KBSEE and KAPTUR have application in any organization that is responsible for the timely and economic development of large software systems. Additionally, any organization responsible for the sustained engineering of large systems over a long period of time could profit from the design knowledge capture capabilities being investigated.

ACKNOWLEDGEMENTS

The KBSEE system was developed with major support from Dr. Hassan Gomaa, Dr. Larry Kerschberg, Dr. Richard Fairley, Chris Bosch, Vijayan Sugumaran, Iraj Tavakoli, and Elizabeth O'Hara-Schettino of the George mason University.

The KAPTUR system prototype was developed with major support from Dr. Sidney Bailin, Manju Bewtra, and Dick Bentz from CTA, Inc. Mike Moore, formerly of CTA but now with NASA/Goddard, also contributed to the KAPTUR development activity.

The success of the current systems is due to the creativity and hard work of these individuals.

REFERENCES

1. Gomaa, H., R. Fairley, L. Kerschberg, "An Evolutionary Domain Life Cycle for Software Maintenance", Report for NASA, 1991

2. Gomaa, H., L. Kerschberg, "An Evolutionary Domain Life Cycle for Domain Modeling and Target System Generation', Report for NASA, 1991

3. Bailin, S., R. Gattis, W. Truszkowski, "A Learning-Based Software Engineering Environment for Reusing Design Knowledge", Report for NASA, 1991

Copies af these reports are available from Walt Truszkowski

N92-22471

REDUCING THE COMPLEXITY OF THE SOFTWARE DESIGN PROCESS WITH OBJECT-ORIENTED DESIGN

M. P. Schuler
(804) 864-6732
NASA Langley Research Center
Hampton, VA 23665-5225

ABSTRACT

Designing software is a complex process. The purpose of this paper is to describe and illustrate how Object-Oriented Design (OOD), coupled with formalized documentation and tailored object diagraming techniques, can reduce the complexity of the software design process. The OOD methodology described uses a hierarchical decomposition approach in which parent objects are decomposed into layers of lower level child objects. A method of tracking the assignment of requirements to design components is also included. Increases in the reusability, portability and maintainability of the resulting products will also be discussed. This method was built on a combination of existing technology, teaching experience, consulting experience, and feedback from design method users [1] [3]. The concepts discussed in this paper are applicable to hierarchal OOD processes in general. Emphasis will be placed on improving the design process by documenting the details of the procedures involved and incorporating improvements into those procedures as they are developed.

INTRODUCTION

A simplified version of an actual project design, for a distributed dynamic controls system, will be used as a case study in describing the OOD process. The controls system was required to: obtain inputs from analog sensor devices attached to a large structure; convert those inputs into digital form; calculate actuator output commands based on the sensor inputs; perform a digital to analog conversion on the actuator commands and send those analog commands to actuators connected to the structure. The intended outcome of this closed loop process was to control the structures movement However, the design examples used for illustration will primarily be concerned with the subsystem responsible for system configuration and data recording, since it does not require a detailed understanding of the application domain. Figure 1 defines the design symbols which will be used in the examples.

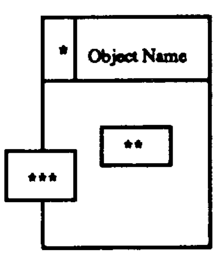
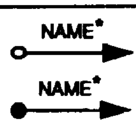
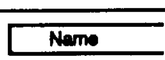
	This symbol represents an object. * This rectangle is inserted if the object is either a terminal or external object. A terminal object is an object that cannot be decomposed any further since each requirement assigned to it is executed by the object itself or its operations. If the object is terminal, a T is placed in this rectangle. An external object is an object which is allocated to a physically separate subsystem and is represented by placing an E in this rectangle. ** Inside this rectangle is the name of an operation that is internal to the object. *** Inside this rectangle is the list of callable operations on the object.
	'Data Couple' is the technical term used for this symbol. *If the circle of the tail of the arrow is empty, the NAME represents the data flow associated with the data couple. If the circle is solid, it represents the command associated with the data couple.
	This symbol is a control arrow. The object at the tail of the arrow controls the object at the head of the arrow.
	This symbol represents a data source or a data sink.

Figure 1. Basic symbol definitions.

This paper focuses on the preliminary design phase. It is assumed that prior to this phase a thorough requirements analysis has been performed and a software requirements document has been completed. The analysis results and the software requirements document are the input documents to the preliminary design phase.

An important goal from the start of this design was to partition the modules of the support domains from those of the application domain (the dynamics controls domain). In other words, to design the system so that code modules produced for different domains would be loosely coupled. This would reduce the complexity of the design and also produce products that were highly reusable, portable, and maintainable.

PRELIMINARY DESIGN DECOMPOSITION STEPS

During the preliminary design phase, a step-by-step process for object identification and decomposition was applied iteratively. For discussion purposes, the object being decomposed will be called a parent object. The objects it is decomposed into will be called the child objects. The following provides a description of each of the steps (Figure 2).

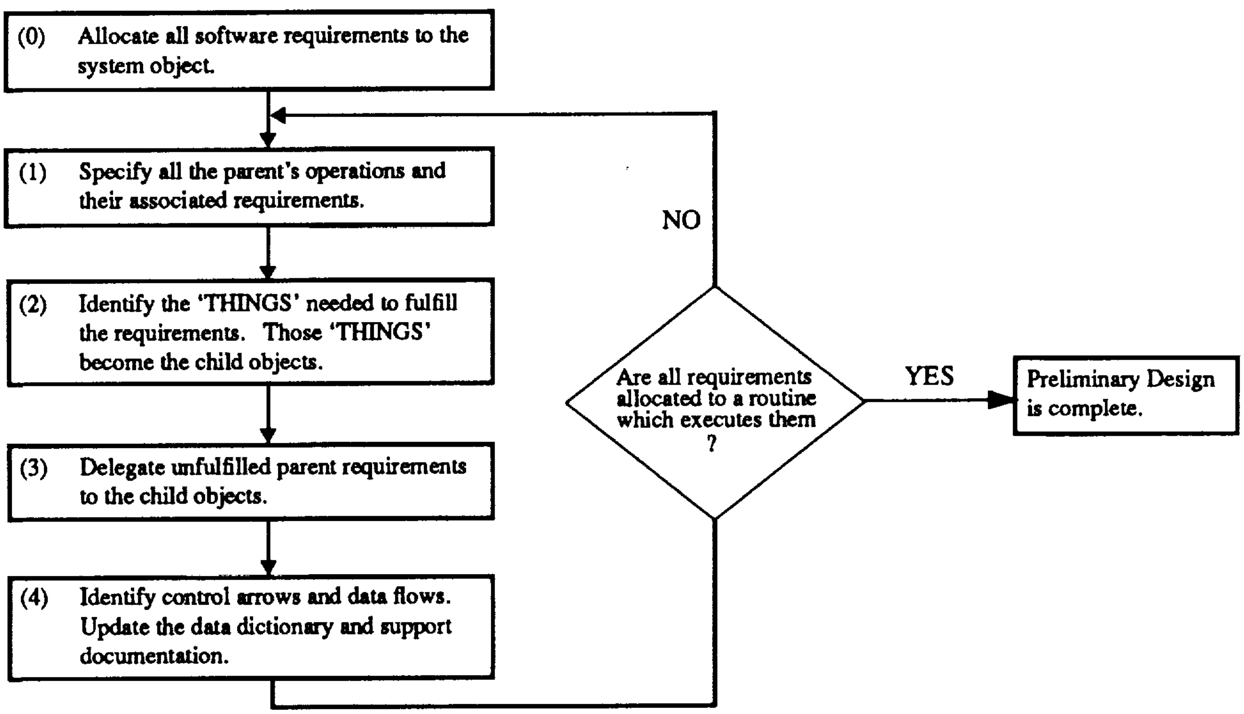


Figure 2. Preliminary design decomposition steps.

(0) Allocate all software requirements to the system object. This is the parent object to the first level of the system's decomposition. This step is done only once.

(1) Review the parent object's requirements and identify all the parent's operations. All requirements must be fulfilled by the parent itself or assigned to one of the parent object's operations. However, those operations may not execute all of the requirements, since during decomposition many of the requirements may be delegated to, and fulfilled by, the child objects or their operations. A textual description, along with the allocated requirements for each of the operations, is recorded on an Operations Description form (Figure 3).

(2) Identify all the child objects. Step through the parent object's requirements, in the order in which they would be executed. The purpose is to determine the primary THINGS needed to fulfill the requirements. In other words,

walk through what needs to be done to determine what THINGS are needed to do it. For each of those THINGS a child object is created thus defining the parent object's decomposition. A textual description of each child object is recorded in the Object Description forms (Figure 3).

(3) All unfulfilled requirements from the parent object are decomposed and assigned to the child objects. All assignments are recorded in the Object Description forms.

(4) Control arrows and data flows between objects are identified and diagramed. A data dictionary is updated and an Object or Operation Description is completed for each element of the design.(Specifying the detailed control and data flow between objects helps identify operations as well as reduce the subjective nature of the object design diagrams.)

OBJECT DESCRIPTION

NAME:
Specify the object name and library number.

VERSION NUMBER / DATE:
This number and date is updated each time the description is updated.

DESCRIPTION:
A brief written description of what the object is required to do.

REQUIREMENTS:
Specify the requirements allocated to this object.

OPERATIONS AND PARAMETERS:
Callable operations on this object.

ASSUMPTIONS:
List assumptions made concerning those things needed to fulfill this objects requirements.

INTERNAL INFORMATION:
Specify internal objects and operations.

ISSUES:
Unknowns that must be determined before this object description can be considered complete.

OPERATION DESCRIPTION

NAME:
Specify the operation name and library number.

OJBECT:
Specify parent object name and library number.

VERSION NUMBER / DATE:
This number and date is updated each time the description is updated.

DESCRIPTION:
A brief written description of what the operation is required to do.

REQUIREMENTS:
Specify the requirements allocated to this operation.

PARAMETERS AND TYPE:
Specify the parameters and types if known.

EXCEPTIONS:
List all exceptions identified thus far.

ASSUMPTIONS:
List assumptions made concerning those things needed to fulfill this operations requirements.

ALGORITHM:
Give the algorithm/pseudocode specifying what the operation will do to fulfill its requirements.

Figure 3. Object and Operation Description Forms.

Steps 1 through 4 are repeated until all system requirements have been allocated to an object or operation which executes them. Requirements allocation is a two-step process. In step 1, all the requirements not executed by the parent are allocated to the parent's operations. In step 3, requirements that were not fulfilled by the parent or its operations are decomposed and allocated to the child objects. If a child object is not terminal[1], it then becomes a parent object and is decomposed. To assure that each requirement is executed by some part of the design, a requirements traceability matrix is constructed. The matrix traces the correspondence between the requirements and

1. A terminal object is an object that cannot be decomposed any further since each requirement assigned to it is executed by the object itself or its operations.

the objects or operations that execute them. Assuring the traceability of requirements to the design is achieved by verifying that: all requirements are listed in the matrix; that an object or operation is assigned to fulfill each requirement; and that those requirements are specified in the description forms.

PRELIMINARY DESIGN

Dynamic Controls System Object Decomposition

The first object to be defined in the preliminary design was the Dynamic_Controls_System (Figure 4), which represented the system in its entirety. All software requirements were delegated to this system object. It was then decomposed into three child objects, one for each of the computer subsystems specified in the requirements. The System_Manager was one of the three child objects defined at this level. The other two child objects will be referred to as Subsystem_One and Subsystem_Two. An Object Description form was drafted for each of the objects defined thus far. The form was used to capture all the available information about an object and therefore included a detailed textual description of this level of decomposition (Figure 3). A brief description defining what each object is required to do was included. All the system requirements were broken down and assigned to the three child objects. These assignments were also recorded in the Object Description forms.

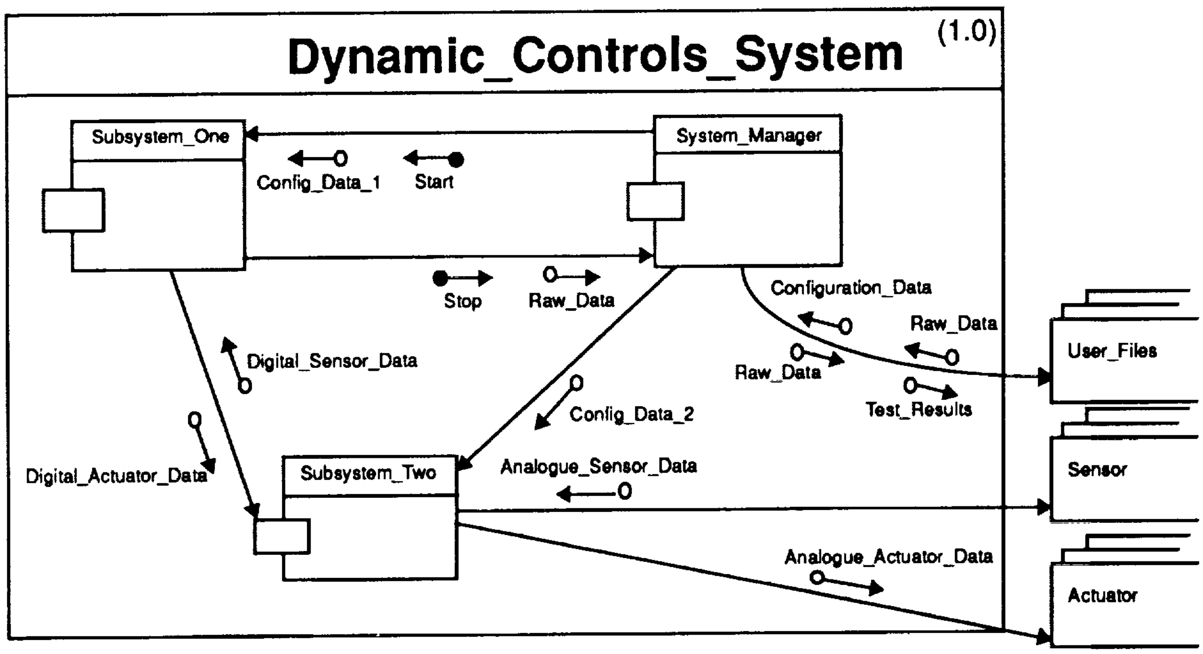


Figure 4. The parent object Dynamic_Controls_System is decomposed into 3 child objects: System_Manager, Subsystem_One, and Subsystem_Two. In the delivered system, all communications between the 3 objects were conducted over a MIL-STD bus.

Object control and communications were defined by diagraming the control arrows and data flows. All the data flows were logged in a data dictionary. A definition was written for each data entry and any applicable requirements were also referenced. Notice that there are control arrows pointing in both directions between Subsystem_One and System_Manager. When the system is being configured, System_Manager is in control. After configuration, Subsystem_One assumes control of both of the other objects. This type of information, which is not recorded on the diagrams, is logged in the Object Descriptions for each of the objects involved. For example, state transition can be recorded in a state transition table and references to that table can be included in the Object Descriptions for each of the objects affected. The command which causes a state to change can be diagramed using data couples as shown by the Start and Stop commands in Figure 4.

Once the requirements, control arrows, and data flows had been specified it was possible to identify the operations on the child objects. For each operation identified, an Operation Description was drafted (Figure 3). A brief description of what the operation was required to perform was recorded. The object's requirements were then assigned to specific operations and those assignments were logged in the appropriate Operation Description form. It is important to note that, all the operations on the objects and all the inputs and outputs to the objects had been thoroughly documented both graphically and textually with the use of the object diagrams and the description forms. Therefore, each object had a clearly defined interface. By first assigning all the system requirements to the three child objects, and then thoroughly defining the interfaces between those child objects, the complexity of the remaining design decomposition was considerably reduced. It was then possible to concentrate on the decomposition of a particular child object, and its requirements, to the exclusion of all others.

System_Manager Object Decomposition

The first level of decomposition was very straightforward since there was a one-to-one correspondence between the computer subsystems and the first level of child objects. However, the decomposition of the System_Manager was not as straightforward. Far too many objects had been identified for a single layer of decomposition and there was no apparent way of grouping them into a logical hierarchy. (A goal of seven, plus or minus two, objects per level of decomposition had been established to minimize the complexity of the design.) The System_Manager had three states of operation; configure the system for a test, record raw data during the test, and post process the raw data. To reduce the complexity of System_Manager it was decomposed into three state manager objects; Pre_Test_Manager, Test_Manager and Post_Test_Manager (Figure 5). Part of System_Manager's requirements were delegated to the internal operation Execute, which scheduled state transitions by making the appropriate calls on the state manager objects. After all the operations had been defined and documented, the remaining requirements for the System_Manager were then decomposed and allocated to the three child objects. For each, an Object Description was written in which the requirements allocations were recorded. All control arrows and data flows were then diagramed and the data dictionary was updated. All operations on the child objects were identified and their Operation Descriptions were completed. These graphical and textural descriptions thoroughly defined each object's interface. It is important to restate that, the number of objects required to define System_Manager were reduced by breaking the requirements into logical groupings (by state) and using state manager objects to encapsulate those groupings. As a result, the design was partitioned in a way that made it possible to concentrate on the decomposition of a particular state manager object, to the exclusion of the others.

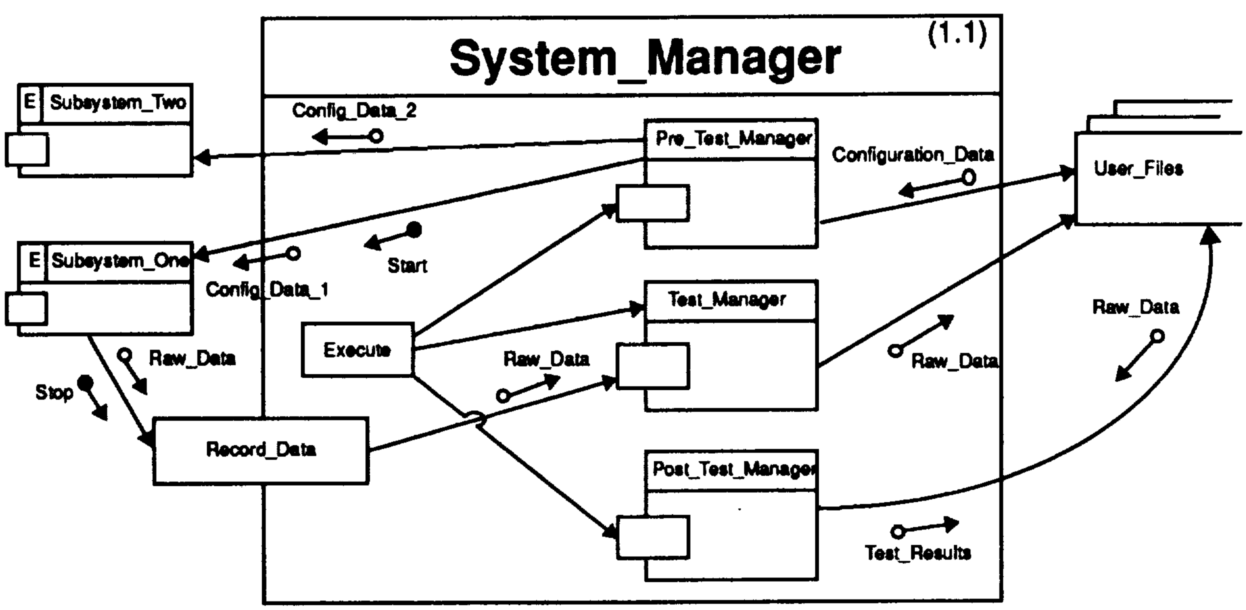


Figure 5. System_Manager is decomposed into 3 child objects: Pre_Test_Manager, Test_Manager, and Post_Test_Manager. System_Manager has one internal operation, Execute.

Pre_Test_Manager Object Decomposition

Pre_Test_Manager was the first state manager to be decomposed. Stepping through the requirements, in the order in which they would be performed, revealed which objects would reside on this level of the design. The first executable requirement of the Pre_Test_Manager was to obtain data for configuring the system. The configuration data was kept on three user-supplied files. These were the THINGS that were needed to fulfill the requirements. Therefore, a child object was created for each of those files; Script_File, Control_File and System_File (Figure 6). These file objects would provide, to Pre_Test_Manager, operations for obtaining the required information. In this way the details of how the configuration information was obtained and file manipulation achieved was hidden from Pre_Test_Manager by the three file objects. Therefore, Pre_Test_Manager could simply make a call on the file objects to satisfy the requirement (obtain data for configuring the system).

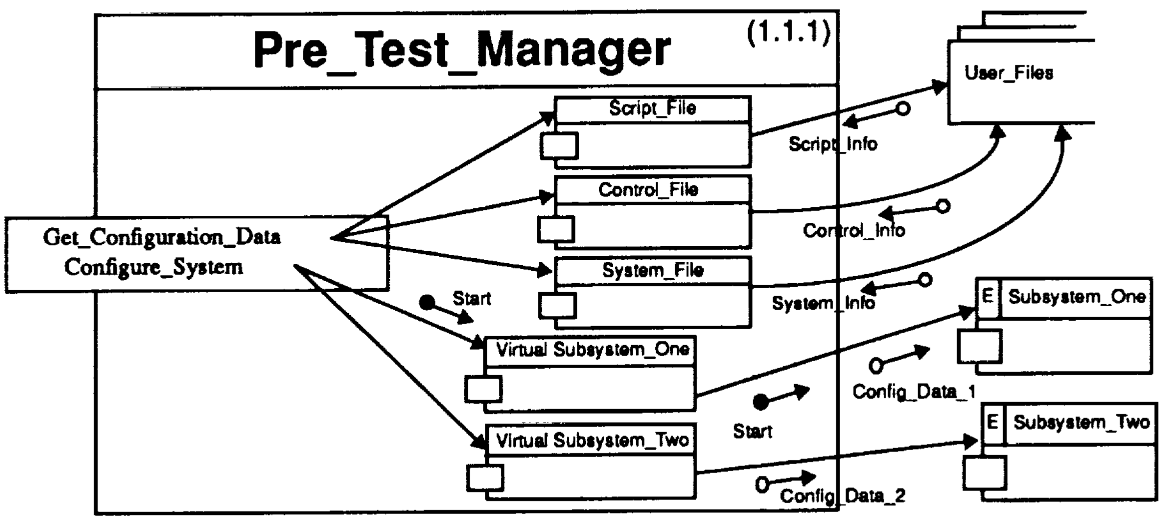


Figure 6. The Pre_Test_Manager is decomposed into five child objects, two of which are virtual objects. Note that the 'E' in the upper left corner of an object designates it as an external object.

Pre_Test_Manager's second requirement was to configure the subsystems with the user-supplied data. However, this was a distributed system and all communications between the System_Manager, Subsystem_One and Subsystem_Two were transmitted through a MIL-STD bus. A bus object was needed to communicate to the other two subsystems. But it was inappropriate to include a bus object at this level of the design, since a strong coupling between bus-related objects and application-related objects at this level of decomposition would substantially reduce the portability and reusability of the resulting components. Therefore, a virtual object[2] was created for both subsystems (Figure 6) [2]. Virtual_Subsystem_One and Virtual_Subsystem_Two would provide, to Pre_Test_Manager, operations to configure the system. Therefore, the complexity of Pre_Test_Manager's decomposition was further simplified by using virtual objects which encapsulated the details of bus communications.

Script_File Object Decomposition

For this case study, assume Script_File had only one operation, Obtain_Script_Data (Figure 7), and all of Script_File's requirements were allocated to that operation. Stepping through those requirements in the order in which they would be executed revealed that opening a file would be the first requirement executed. Therefore a child object, File_Manager, was created. The File_Manager was allocated the requirements for opening the files and handling errors which occurred in that process. As execution continued information would be taken off the file and

2. A virtual object is a logical construct used to represent an external object that resides on a separate processor. The virtual object imitates the external object's interface. An external object is an object which is allocated to a separate subsystem.

put in storage for later use in configuring the system. To do this the child objects, Sensor and Actuator, were created to store information relating to the system sensors and actuators.

Collectively, the three file objects; Script_File, Control_File and System_File provided a partition between the dynamic controls domain and the file management domain. That resulted in a decoupling of the domains. Therefore, the system was more maintainable since changes to the controls domain would not affect file objects and changes to the file system would only effect the file objects and their encapsulated child objects. For example, if a requirements change specified that the actual script file was to be obtained from a network node instead of a file on the disk, the File_Manager object could simply be replaced with a Network_Manager object. Since the Script_File encapsulates all the design elements used to implement input operations, Pre_Test_Manager would be unaffected. In addition, portability was increased since File_Manager was designed to provide general operations having to do with file access so that it could easily be reused. Not only was it reused by the Control_File, System_File and objects in Test_Manager and Post_Test_Manager but it could be reused by other systems in other domains which require disk file access.

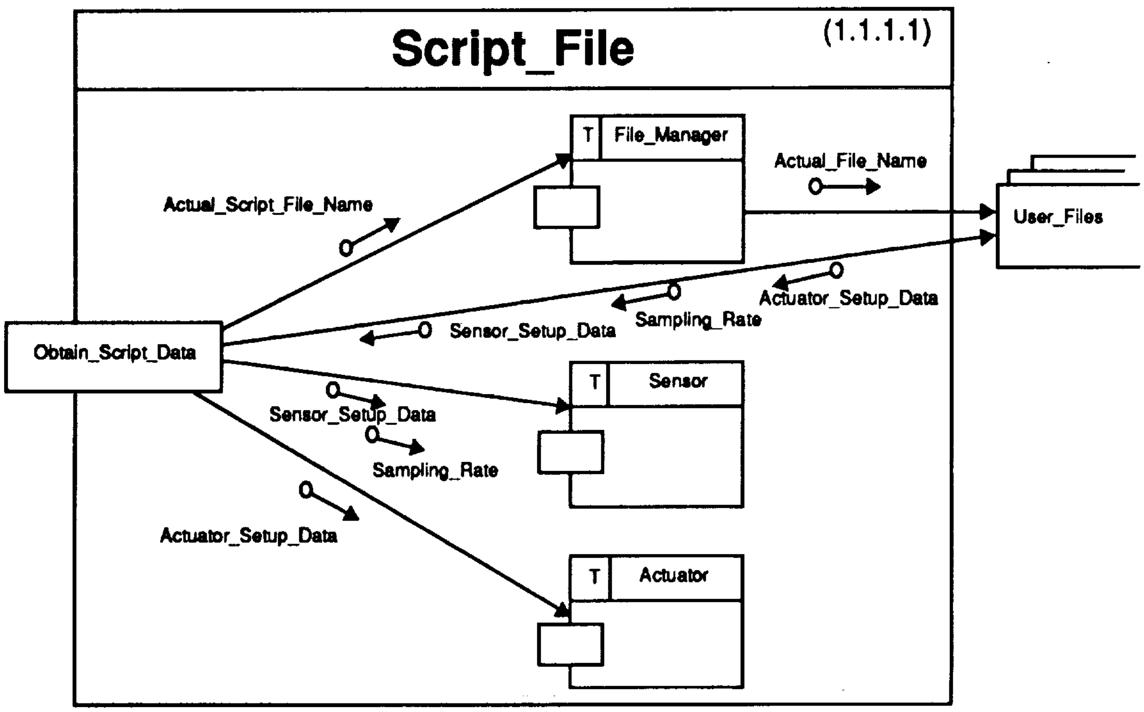


Figure 7. Script_File is decomposed into 3 child objects: File_Manager, Sensor and Actuator. The 'T' in the upper left-hand corner of the child objects indicates that they are 'Terminal Objects'; objects that can not be decomposed any further since all requirements allocated to them are executed by one of their operations.

Virtual Subsystem One Object Decomposition

To perform the operation Configure, Virtual_Subsystem_One needed to access the Sensor and Actuator objects to obtain the information necessary for configuration (Figure 8). That information had been placed in the Sensor and Actuator objects by the three file objects; Script_File, Control_File and System_File. To transmit that information to Subsystem_One, a Bus_Manager was created to encapsulate the details of the communications domain. Since bus management would require complex hardware specific code, it was decided that two separate design efforts would be conducted in parallel: first, the application-level design which dealt with the real world dynamic controls domain; and second, the design of the communications drivers for the MIL-STD bus. The communications driver design was done bottom up, from the card level. Together, figures 8 and 9 graphically show how the two designs were merged. The top level object from the bus design was Bus_Manager. It provided, for example, 'get' and 'put' operations to Virtual_Subsystem_One. In the same manner Virtual_Subsystem_Two reused the Bus_Manager to communicate with Subsystem_Two.

Portability was substantially increased by creating a hierarchical design in which virtual objects were used to partition the application domain components from the bus domain components. For example, controls domain components could be ported to other systems having different communications devices. In addition, the bus communication components could be used to control bus traffic for any application using the same MIL-STD bus and card. Over four thousand lines of code from Bus_Manager have already been reused on another project, and no modifications were necessary even though the application domain was completely different. This was possible since Bus_Manager provided general purpose operations to implement the MIL-STD bus protocol which had no relation to the application domain.

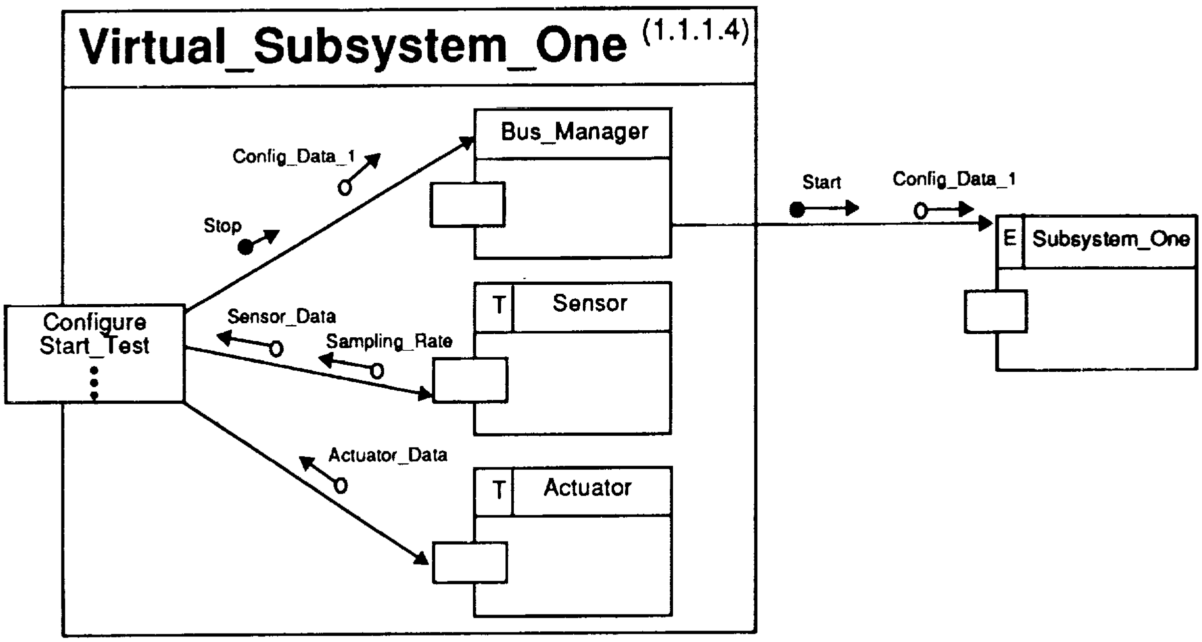


Figure 8. The Virtual_Subsystem_One is decomposed into 3 child objects: Bus_IO, Sensor, and Actuator. Note that Bus_Manager (figure 6) and its child objects facilitate access to the external object Subsystem_One.

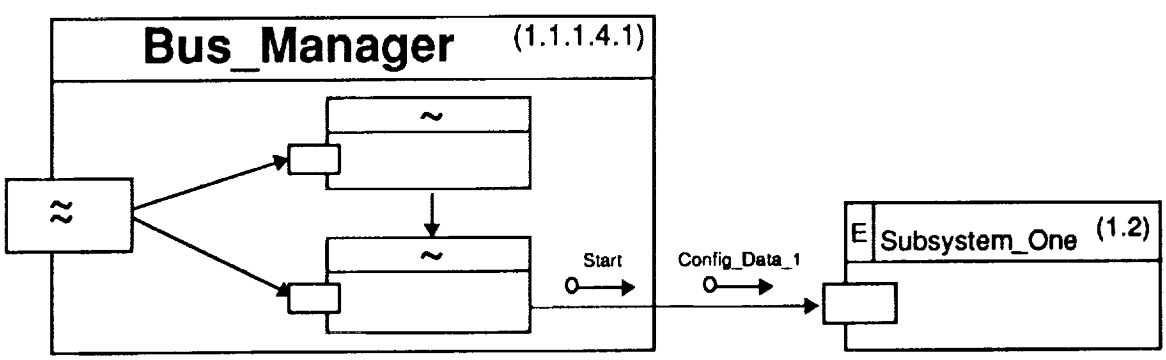


Figure 9. Bus_Manager is decomposed into two child objects which contain routines that control communications across the MIL-STD bus.

Extensions to the Preliminary Design Decomposition Steps

After reviewing the completed preliminary design it was evident that certain generalizations could be made about the decomposition process which could augment the preliminary design decomposition steps shown in Figure 2. These were recorded to provide additional insight into the design process for future projects and to confirm that these concepts worked successfully.

Virtual Objects.

If the THING that is needed to fulfill a requirement is an external object residing on a separate processor:

(A) Create a virtual object to represent that THING and assign to it the operations required (by the parent) to manipulate the external object.
(B) Then create a child object to manage the details involved in controlling the communications device used to access the external object. The communications manager object should provide only those operations specific to the defined protocol for that device.

Support Domain Access.

If services from a support domain, such as file management, are required to access the THING that is needed to fulfill the requirements:

(A) Create an object that will represent that THING and assign it the operations necessary to fulfill the parent requirements.
(B) Create a child object to manage the implementation of the services required by that support domain. This domain manager object should provide only those operations required to manipulate elements under its domain.

Both of these techniques are used to partition the design so that objects related to different aspects of the solution are loosely coupled which increases the portability of the resulting software components. Also, the domain/device manager objects encapsulate implementation details and provide a controlled interface through which services are obtained. This increased the maintainability of the resulting system in two ways: first, any changes related to the domain/device would be localized to the encapsulating object; and second, modifications to other objects would not effect the internal implementation of the domain/device object.

State Managers

If the parent object has several states, and a number of objects associated with each state, a child state manager object should be created for each of the states to reduce the complexity of the remaining design decomposition.

Mixing activities from preliminary and detailed design is one of the most common mistakes designers make. It is important to refrain from considering implementation details or data types until the detailed design phase. During the preliminary design, emphasis should be placed on what objects are necessary to fulfill the requirements, rather than on how requirements could be implemented.

DETAILED DESIGN

The general rules for transitioning from preliminary to detailed design were fairly straightforward. All the objects and operations were converted to Ada Program Design Language (PDL). Each object was made into an Ada package or task. Each operation was made into an Ada function or procedure and the data flows and the data dictionary were used to determine the data types for the operation parameters. Any alterations, additions or deletions in the design were documented by updating the preliminary design documentation. The Object and Operation Description forms from the preliminary design were reused to document the detailed design. The descriptions were copied into the prologues of the packages and operations. Since these descriptions documented the requirements allocated to each preliminary design element, traceability from requirements to detailed design was maintained. Also, the algorithms from the Operation Descriptions were inserted, along with null statements, into the Ada functions and procedures. The design elements were then compiled to verify the Ada interfaces. In addition, the PDL and the code were both done in Ada, so the process of converting the PDL to the completed code was just a matter of coding the algorithms specified within the PDL. Since the code also contained the documentation which specified the allocated requirements, traceability from the requirements to code was also achieved.

DESIGN DOCUMENTATION

A well defined method of documentation is invaluable. It basically eliminated the subjective nature of the preliminary design diagrams. The Object and Operation Description forms (Figure 3) supplement the object diagrams and provided an opportunity for the designers' intentions to be documented. Although it has not been discussed in this paper, library numbers were used to uniquely identify each graphical element of the design. To assure that the proper description was associated with each graphical element, those numbers were also recorded in the description forms [3]. When the preliminary design was completed the diagrams and accompanying description forms contained enough information to implement the detailed design. In addition, the description forms were used to trace the requirements allocation and build the requirements traceability matrix. A Decomposition Tree was also made which pictorially represented the parent/child hierarchy [3]. This was used as a quick reference guide and also as an aid in locating reuse opportunities along different branches of the design. In addition, it can also be used by management to track the progress of the design activity. An accurate representation of the current projects configuration can be maintained by updating these documents during each phase; detailed design, implementation, testing, and delivery. Collectively these documents can serve as the 'As Built Configuration Document' which describes how the functional specifications were achieved in the final product.

PROCESS IMPROVEMENT

Many strides were made in OOD process improvement during this project. The most significant of these was to clearly define the process itself. Figure 10 shows a graphical representation of the process. By determining the process, as well as the steps and procedures followed at each phase, a baseline for process improvement is defined. As future projects reuse this process, procedural improvements can be added to the baseline and the list of lessons learned can be augmented. This information can also be exchanged with other organizations using similar methods. To facilitate this, an individual in each organization is given the responsibility of recording the current state of the process, discovered improvements, and lessons learned. Not only are improvements and lessons learned recorded, but an attempt is made at documenting the rationale behind them. Each organization is responsible for feeding this information back to a central person, the 'keeper of the method.' This person is responsible for collecting from each organization the improvements and rationale, updating the method accordingly, and redistributing it to all the organizations involved. So far, these organizations include two NASA centers, ESA, and several commercial companies. Although this network is in its infancy, it is spreading nationally as well as internationally.

CONCLUSION

With the OOD procedures outlined in this paper, the complexity of the preliminary and detailed design process can be substantially decreased. In addition, the reusability, portability, and maintainability of the resulting products will be increased. Also, process improvement can be obtained by documenting the details of the procedures involved and incorporating successfully demonstrated improvements into those procedures.

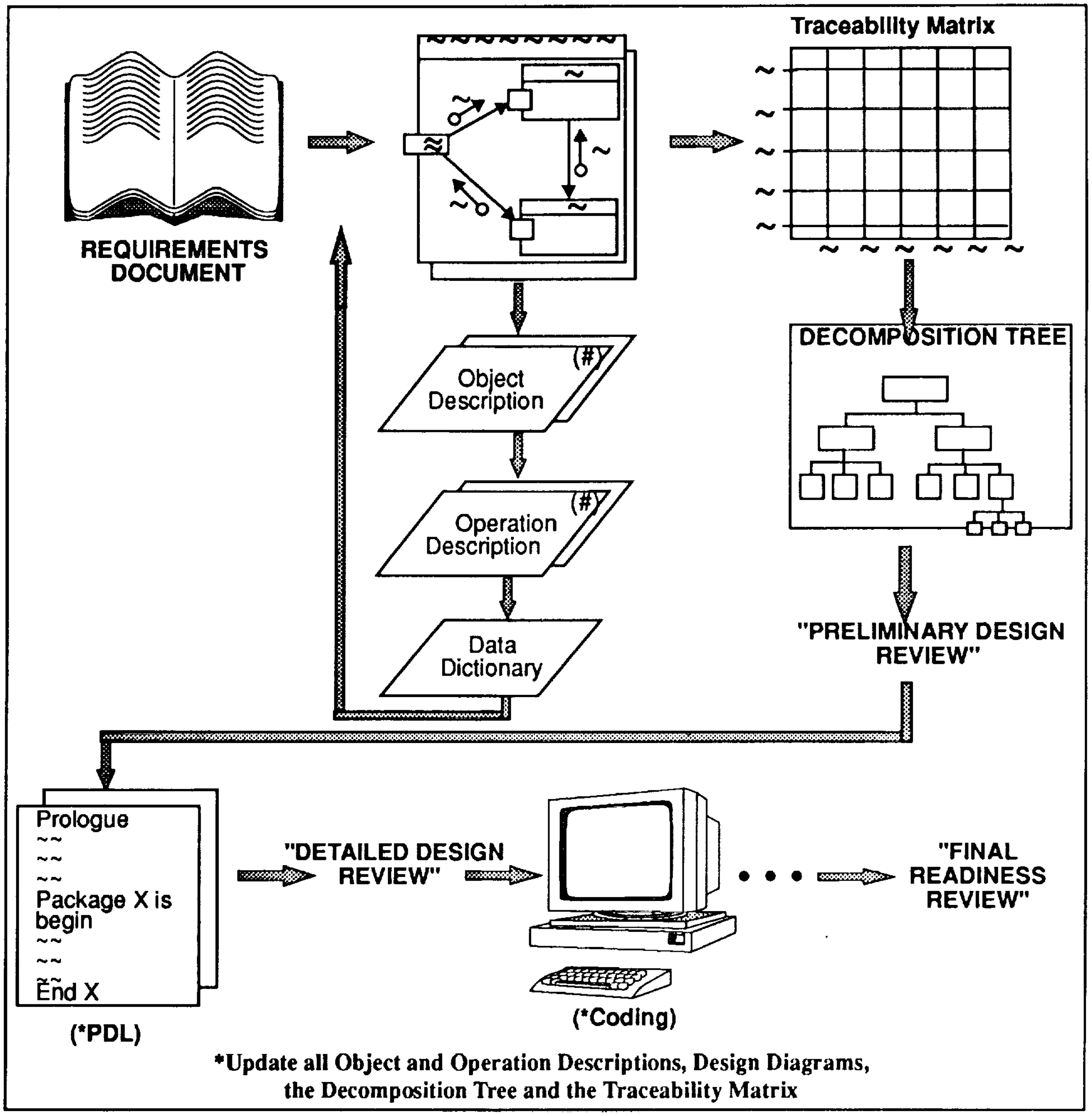


Figure 10. Process diagram.

REFERENCES

1. [Anderson 91] Anderson,J.,et al., Manageable Object-Oriented Development: Abstraction, Decomposition, and Modeling, Proceedings of Tri-Ada'91, San Jose, CA., October 21-25, 1991.

2. [Mc Quown 89] McQuown,K.L. Object Oriented Design In A Real-Time Multiprocessor Environment, Proceedings of Tri-Ada '89, Pittsburgh, PA., October 23-26, 1989, pp. 570-588.

3. [Schuler 91] Schuler,M.P.,Evolving Object Oriented Design, a Case Study, Proceedings of the Eighth Washington Ada Symposium (McLean, VA., June 17-21, 1991), pp.50-61.

DATA AND INFORMATION MANAGEMENT

(Session C1/Room C4)

Wednesday December 4, 1991

- **Techniques for Efficient Data Storage, Access, and Transfer**
- **A Vector-Product Information Retrieval System Adapted to Heterogeneous, Distributed Computing Environments**
- **AutoClass: An Automatic Classification System**
- **Silvabase: A Flexible Data File Management System**

DATA AND INFORMATION MANAGEMENT

N92-22472

ADVANCED TECHNIQUES AND TECHNOLOGY FOR EFFICIENT DATA STORAGE, ACCESS AND TRANSFER

Robert F. Rice
Jet Propulsion Laboratory
Califronia Institute of Technology
Pasadena, CA 91109

Warner Miller
Goddard Space Flight Center
Greenbelt, MD 20771

ABSTRACT

Advanced techniques for efficiently representing most forms of data are being implemented in practical hardware and software form through the joint efforts of three NASA centers. These techniques adapt to local statistical variations to continually provide near optimum code efficiency when representing data without error. Demonstrated in several earlier space applications, these techniques are the basis of initial NASA data compression standards specifications.

Since the techniques clearly apply to most NASA science data, NASA invested in the development of both hardware and software implementations for general use. This investment includes high-speed single-chip VLSI coding and decoding modules as well as machine-transferrable software routines. The hardware chips have been tested in the laboratory at data rates as high as 700 Mbits/s.

A coding module's definition includes a predictive preprocessing stage and a powerful adaptive coding stage. The function of the preprocessor is to optimally process incoming data into a standard form data source that the second stage can handle. The built-in preprocessor of the VLSI coder chips is ideal for high-speed sampled data applications such as imaging and high-quality audio, but additionally, the second stage adaptive coder can be used separately with any source that can be externally preprocessed into the "standard form." This generic functionality assures that the applicability of these techniques and their recent high-speed implementations should be equally broad outside of NASA.

INTRODUCTION

Science data returned from space instruments is often studied in great detail by many investigators. For some investigators the precise value of individual data samples can be crucial. Such "high fidelity criteria" led NASA to the development of efficient "lossless" techniques for representing such data without introducing error.

Increasing data rate requirements of many new instruments and a demonstrated capability to support a broad range of instrument data fostered the recent implementations of an important subset of these techniques as single coding and decoding modules. A somewhat broader set of algorithms is being incorporated into a software package written in C.

The intent of this paper is to introduce the underlying characteristics and algorithms associated with this hardware and software at a tutorial level. Details are provided by References 1-4.

The Standard Source

By various means, many data sources can be converted to one with the basic characteristics in Figure 1. Inactive sources will generate a greater occurrence of small data values than active sources. Conversely, active sources will generate a greater occurrence of larger data values than inactive sources. But in both cases (and all the cases inbetween), smaller data values occur more frequently than large.

Many real data sources can be preprocessed into the form described for Standard Sources.[1] So in subsequent discussions we will presume that this step has been done unless noted otherwise. Eventually, we will return to the preprocessing step.

The Variable Length Code

A codeword is a unique sequence of binary bits used to represent data values from this Standard Source. All the codewords together make up a "code." Basic data systems use a fixed-length code. That is, suppose there were 2^n data values. Then the fixed length code would consist of 2^n codewords, all n bits in length. Thus every data sample would require n bits/sample. By using a variable length we can usually improve on this by taking advantage of the differing frequency of occurrence shown in Figure 1.

A variable length code has different codeword lengths, as shown by the example in Table 1.

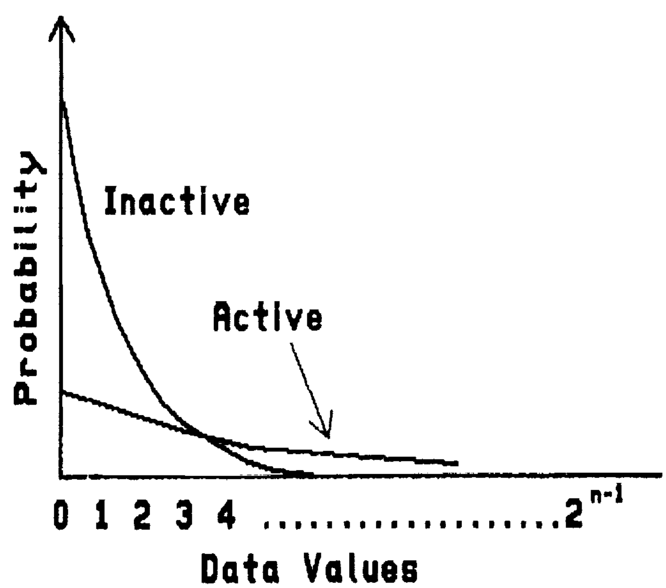


Figure 1. Standard Source Characteristics

---

[1]This preprocessing should also produce uncorrelated data values.

DATA VALUE	CODEWORD
0	1
1	01
2	001
3	0001
4	00001
•	•
•	•
•	•

Table 1. Simple Code

If the small data values occur very frequently, one could expect the average number of bits used to be less than n bits/sample (because the shorter codewords would be used more frequently). For example, if only the data value 0 occurred, the simple code of Table 1 would use only 1 bit/sample. But the opposite is true, too. The code in Table 1 might use a lot more than n bits/sample when used with a very active data source.

The Huffman Code

D. A. Huffman invented an algorithm for generating the best code for a "known" distribution of data values, as shown in Figure 1 [5]. This suggests a possible solution. Can we just use the right Huffman code? Unfortunately, such a direct application of the famous algorithm can have some practical difficulties:

1. Wrong Code: Data activities tend to vary with time, from instrument to instrument, and may not be known a priori at all.

2. Implementation Problem: Some new instruments have up to 2^{14} data values (e.g., 14 bits/sample fixed length code). Thus a single Huffman code could conceivably require a lookup table containing 2^{14} codewords, some of them quite lengthy.

The Adaptive Coder

We seek an answer to both of these problems with the structure of Figure 2. Consider functionally what this adaptive coder structure does. It operates on short blocks of data, X (e.g., 16 data samples), choosing the best of N coders to use for each block. A separate identifier, ID, precedes the chosen coded block, $C_{ID}[X]$, to tell a "decoder" which decoding algorithm it needs to use. The identifier penalty is small. For example, a coder with $9 \leq N \leq 16$ code options requires a 4-bit identifier (or 0.25 bits/sample for a 16-sample block). Thus with properly chosen code options, such an adaptive coder should be able to handle large variations in data activity.

This is in fact the case, and the implied complexity never materializes. The chosen codes are both equivalent to Huffman codes but require no table lookups at all.

ADAPTIVE VARIABLE-LENGTH CODER CHARACTERISTICS

References 1 and 3 provide several variations to the adaptive variable length coders (AVLC) that begin with the structure in Figure 2 as their starting point. We will focus here on the most useful variations which have recently been implemented as individual VLSI chips. The reader should consult the references for greater detail and a broader perspective.

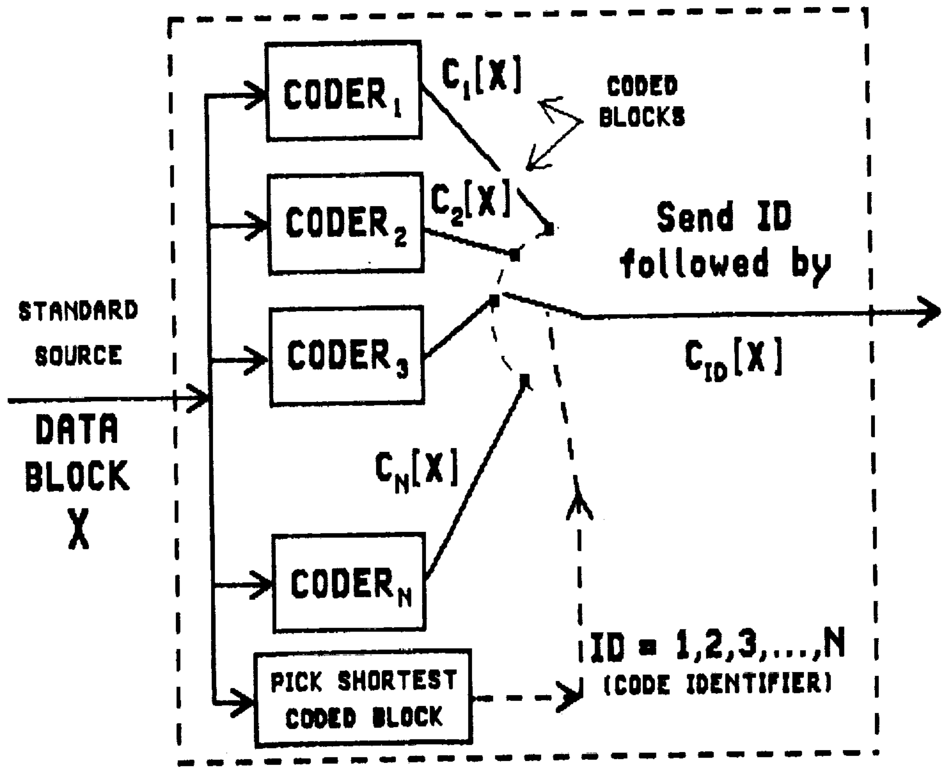


Figure 2. Adaptive Variable Length Coder Concept

Entropy

Information theory provides us with a mathematical measure of coding efficiency. For the case discussed here, entropy is a function of the probability distributions in Figure 1. Basically, entropy increases with activity. More specifically, a Standard Source with a fixed data value distribution and associated entropy, H bits/sample, cannot be coded with fewer bits/sample than H. A coder which codes close to the entropy, such as a Huffman code used on its design distribution, is said to be efficient.

AVLC Options

The simplest variable length code is the one shown in Table 1. It is defined for any number of data values. A codeword for data value j is j zeroes followed by a one. Clearly this code requires no table lookup. Further, it is an efficient coder over the entropy range $1.5 \leq H \leq 2.5$ bits/sample.

Other code options first split an n-bit data sample into its k least significant bits and its n-k most significant bits. Then the simple code of Table 1 is applied to the samples formed by the most significant bits.. Each value of k provides a code option which is efficient over an entropy range of about 1 bit/sample, centered on an entropy of k +2 bits/sample. More amazingly, it has been shown that these simple "Split-Sample" code options are equivalent to Huffman codes.[2]

Thus with enough of these easily implemented code options, the average performance of the adaptive coder in Figure 2 will look like that shown in Figure 3. That is, such an AVLC will be "efficient" everywhere except at very low entropies.

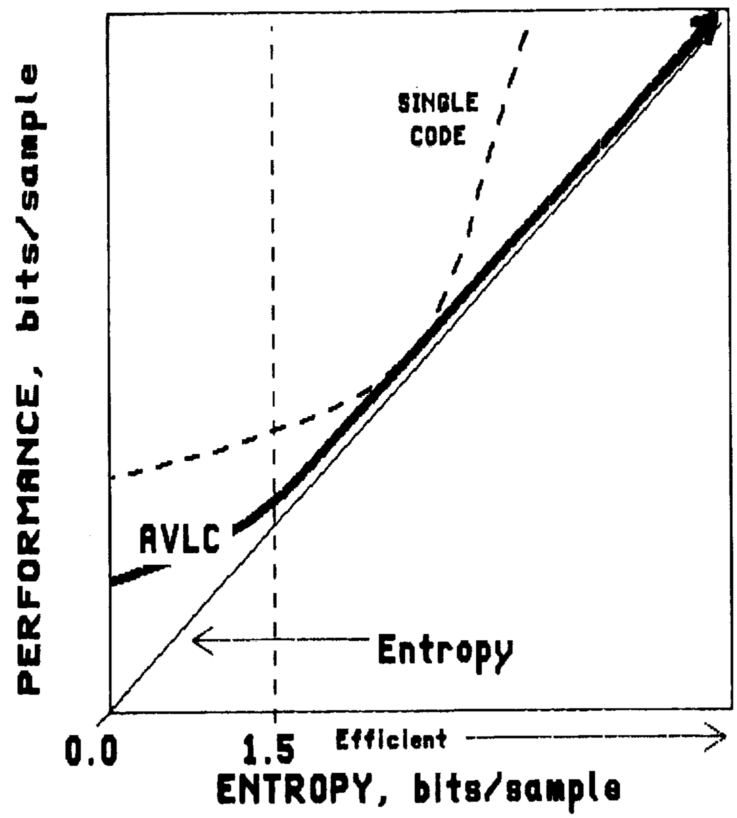


Figure 3. AVLC Performance Characteristics

CODING MODULE

For a vast number of real problems, the desired step to accomplish the conversion of an instrument data source into the desired standard source form can be accomplished by a simple predictive preprocessor. We then define an overall "coding module" in Figure 4 that incorporates this simple preprocessor along with an AVLC.

First note that a coding module includes a switch with positions C and D. In position C, the AVLC sees the output of the "built-in" preprocessor whereas in position D, the AVLC becomes available for directly coding the output of any external preprocessor.

Now consider the built-in preprocessor itself. For the moment assume that the switch with positions A and B is in position A. Then the output of the "sample delay" is the previous input sample. This acts as a sample prediction that the next sample equals the last. The result of differencing, Δ, is then the error in this prediction. For many problems this is a very good prediction and little can be gained by greater sophistication. But just in case, switch position B allows the module to use an arbitrary external prediction.

In either case, errors tend to be distributed around zero. That is, small differences are generally more likely than larger differences. The MAP function simply converts the differences into the integers expected for a Standard Source (0 maps to 0, -1 maps to 1, +1 maps to 2, -2 maps to 3, +2 maps to 4, and so on). See Refs. 1 and 3 for subtleties on this mapping.

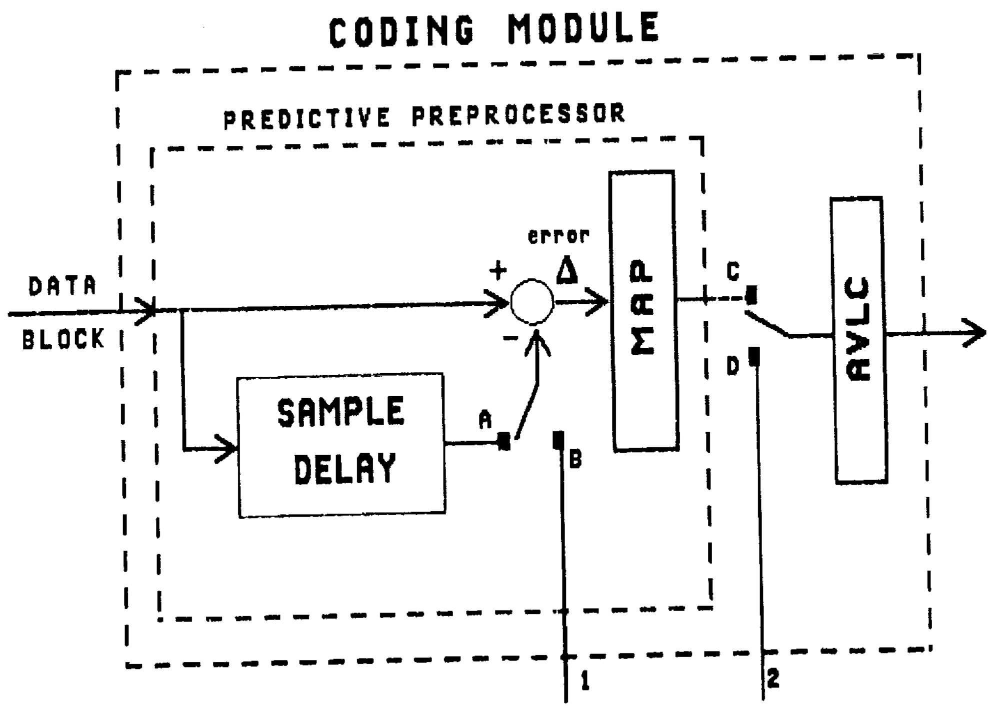


Figure 4. Coding Module

RECENT IMPLEMENTATIONS

Hardware

Two separate programs were recently completed to implement similar versions of the coding module of Figure 4 in high-speed, low-power CMOS VLSI.

The Jet Propulsion Laboratory (JPL) focussed on the design and fabrication of both gate array and standard cell coding module chips. Both were successfully tested in the laboratory in September 1990 at data rates up to 180 Mbits/s.

The JPL developments were driven by severe schedule constraints to meet the specific requirements of a then on-going project. These chips can operate on data quantized to 12 bits/sample (2^{12} data values) and incorporate an 11-option AVLC. The CRAF/Cassini project is currently funding a second-generation chip which will become flight qualified for these missions.

The University of Idaho, under the direction of Goddard Space Flight Center (GSFC) focussed on a more generic chip development unencumbered by the burdens of flight project schedules. A coding module and a companion "decoding" module chip set were recently tested in the laboratory at data rates up to 700 Mbits/s. These chips can operate on data quantized from 4 to 14 bits/sample and incorporate a 12-option AVLC. Both switches in Figure 4 are included in the design.

While not the most general coding module (see Ref. 3) this design includes the most important features needed to support a very broad range of problems. Consequently, its algorithmic specification is the basis of initial NASA data compression standards efforts. A second-generation chip set (with some slight enhancements) is planned to provide this capability in space qualified form.

Software

The most general parameterized coding/decoding algorithms [3], of which the University of Idaho/GSFC chip set is a subset, have been implemented in C on a specific computer. Planned refinements to make the software more portable and callable will be made available to potential users.

ACKNOWLEDGEMENT

The research described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology, the Goddard Space Flight Center and the NASA Space Engineering Research Center for VLSI Design, University of Idaho, under a contract with the National Aeronautics and Space Administration.

REFERENCES

1. Robert F. Rice, Pen-shu Yeh, Warner Miller, "Algorithms for a Very High Speed Universal Noiseless Coding Module," JPL Publication 91-1, Jet Propulsion Laboratory, Pasadena, California, February 15, 1991.

2. Pen-shu Yeh, Robert F. Rice, Warner Miller, "On the Optimality of Code Options for a Universal Noiseless Coder," JPL Publication 91-2, Jet Propulsion Laboratory, Pasadena, California, February 15, 1991.

3. Robert F. Rice, "Practical Universal Noiseless Coding Techniques, Part III" JPL Publication 91-3, Jet Propulsion Laboratory, Pasadena, CA, (to be published).

4. J. Venbrux, et al, "A Very High Speed Lossless Compression/Decompression Chip Set," JPL Publication 91-13, Jet Propulsion Laboratory, Pasadena, California, July 15, 1991.

5. D. A. Huffman, "A Method for the Construction of Minimum Redundancy Codes," Proc. IRE, Vol. 40, pp. 1098-1101, 1952.

N92-22473

A VECTOR-PRODUCT INFORMATION RETRIEVAL SYSTEM ADAPTED TO HETEROGENEOUS, DISTRIBUTED COMPUTING ENVIRONMENTS

Mark E. Rorvig
Software Technology Branch
NASA Johnson Space Center
Houston, TX 77058

ABSTRACT

Vector-product information retrieval (IR) systems produce retrieval results superior to all other searching methods but presently have no commercial implementations beyond the personal computer environment. (NELS) NASA Electronic Library System, provides a ranked list of the most likely relevant objects in collections in response to a natural language query. Additionally, the system is constructed using standards and tools (i.e., UNIX, X-Windows, Motif, TCP/IP) that permit its operation in organizations that possess many different hosts, workstations and platforms. There are no known commercial equivalents to this product at this time. The product has applications in all corporate management environments, particularly those that are information intensive, such as finance, manufacturing, biotechnology, and research and development.

INTRODUCTION

The field of information retrieval (IR) has always advanced unevenly. Even fundamental theoretical insights have occaisonally required a generation or more of developments in electrical engineering to enjoy commercial practice. The NASA Electronic Library System (NELS) is a good example of hardware and software dependent intellectual advances. In this case, two discoveries from the 1960's regarding indexing system performance and retrieval ranking, have been implemented within the context of standard operating systems, network protocols, languages, and display tools. Figure 1 below illustrates the interactions among these electrical and software standards and components.

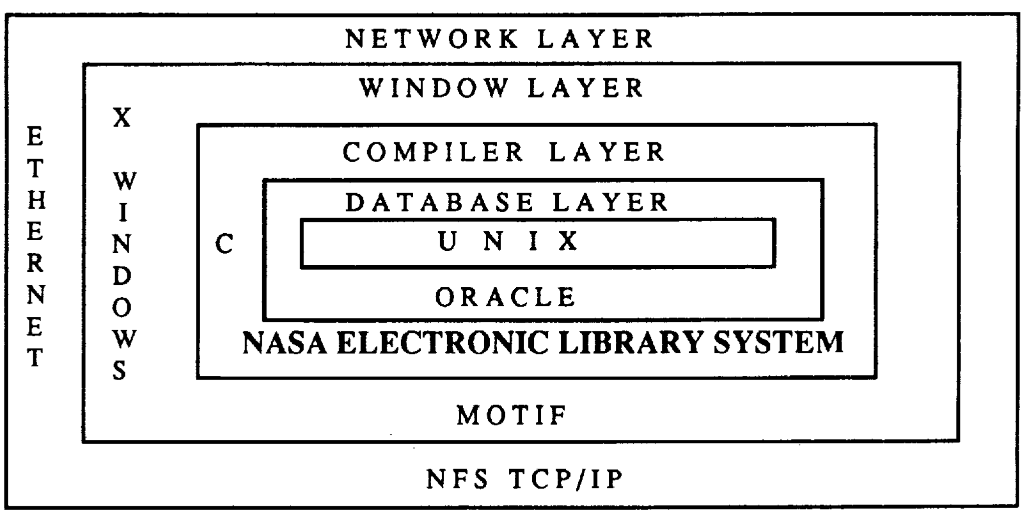


Figure 1: NELS environment variables expressed as layers of interacting electrical and software standards, and their specific implementations. For example, NELS is written in "C", displayed through Motif, shares resources with other devices by NFS TCP/IP, and depends on Oracle and UNIX for system support.

This article will describe the intellectual history of searching supported by NELS 1.0 and the its method of use on wide area networks. A few basic screen layouts will be examined to provide a "feel" of the system for the reader. Flinally, some hardware platforms and configurations currently supported by NELS will be discussed.

VECTOR-PRODUCT SEARCHING AND RELEVANCE FEEDBACK

One of the great counter intuitive discoveries of IR research in the early 1960's was that the precision and depth of indexing did not result in improved retrieval performance when compared to the simple strategy of providing access to objects by every word in the title [1]. This research finding directly influenced the development of the NASA RECON system in the late 1960's, and subsequently the system development of Knight-Ridder's DIALOG, a large commercial distributor of database information. Although the available entry points were expanded to include all words in the abstract of an object, as well as the title and author information, subject indexing diminished in importance in online systems, except as an aid for quicly isolating a specific body of intellectually similar objects for further searching.

This concept has been implemented in NELS in two mutually reinforcing ways. First, although keywords may be assigned by indexers and searched as part of a natural language query, they are not an integral consideration. Rather, objects are assigned to classes, which may consist of intellectual concepts, organizational units, or in NASA's research engineering environment, system, subsystem, and sub-assembly hierarchies. A system user may then navigate to any class body of interest, search further, or simply list the objects. Second, the lexical components of object descriptions in the author, keyword, title, and abstract fields are all parsed, stripped of suffixes, and placed in an object attribute table for searching by natural language.

Searches conducted directly by natural language have long been a cornerstone of IR. However, in nearly all commercially available systems (excluding those dependent on devices such as array processors and their like), searching has been restricted to boolean logic queries; a form difficult to learn, subject to return of null sets in long queries, and in shorter ones, return of objects as an undifferentiated lump of knowledge. NELS has rejected this approach by implementing another important discovery of the 1960's, vector-product searching as defined at Harvard and Cornell by Gerard Salton and his various students and proteges [2].

Salton's method, expressed as the cosine vector approach, was a revolutionary discovery. Implicit within it and years ahead of the field was the first natural language interface. The formula below defines this method, simply stated as the cosine coefficient of commonality between QUERY terms and DOCument terms, or, between a query and all documents sharing at least one term in common with the query.

$$\text{COSINE(QUERYi,DOCj)} = \frac{\sum_{k=1}^{t} (\text{TERMik} \bullet \text{TERMjk})}{\sqrt{\sum_{k=1}^{t} (\text{TERMik})^2 \bullet \sum_{k=1}^{t} (\text{TERMjk})^2}}$$

The result of applying this formula to a query is a list of objects retrieved by their degree of lexical closeness to the query. Further, the degree of closeness desired may be set by a user as a cut-off point associated with the user's sign-on identification. To initiate a search, from anywhere in a class structure, even at the very top, requires of the user only to type a description of his or her information requirement, in as general or specific a form as desired.

Wide implementation of this approach never occurred, though the focus of research in IR itself has indicated conclusively that it is superior to boolean search techniques, regardless of users, data, or system platforms [3]. Though there are many reasons for this phenomenon, chief among them was the basic limitation of memory required to store and search the vector spaces in a timely manner for massive data files. Since the recent arrival of virtual memory and RISC architectures, however, this consideration has diminished significantly in importance and has led NASA to implement this concept in complete and final form.

However, even with relevance information appearing as a cosine "score" beside each item, users may have further difficulty identifying some items and redesigning their query based upon it due to typing errors, misunderstanding of the vocabulary of an abstract of a retrieved document, or inability to learn enough about a document to conveniently restate their queries. Therefore, a system of relevance feedback is further available for users. To use this feature, a searcher pulls down a menu labelled "options" and selects the command "like". A simple click on one of the previously retrieved items with a mouse causes the system to initiate another search, in this case substituting the abstract of the selected item for the earlier natural language query.

By iteratively applying this method, the user is able to bring his query closer and closer into conformance with the language of the system, constantly raising the level of relevance. Indeed, it is possible for a user to find objects with this method with extremely little prior knowledge of the search domain. Figure 2 below illustrates this procedure with a simple flowchart.

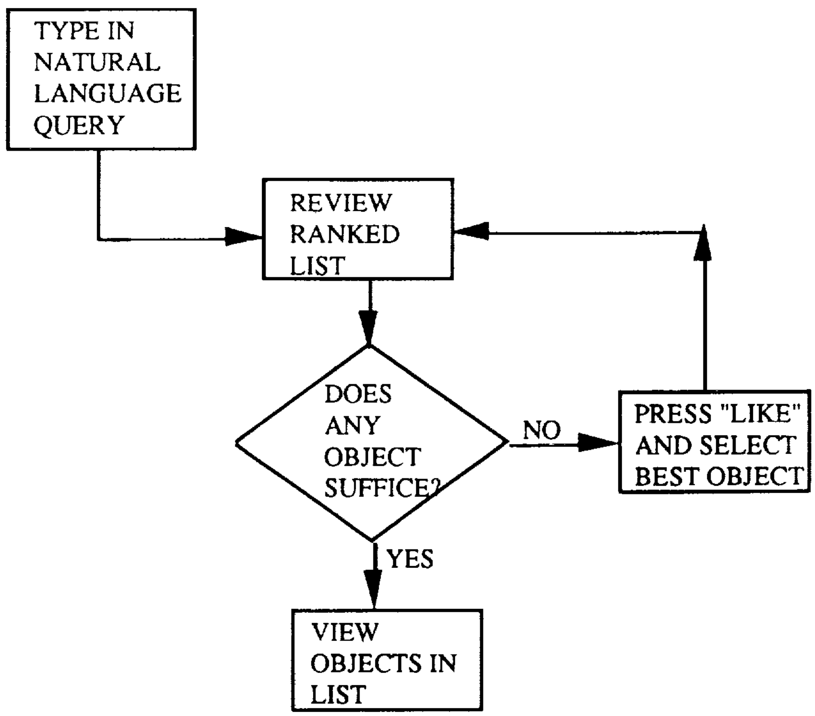


Figure 2: Method for using relevance feedback in NELS to improve system retrieval performance, that is, to move iteratively more closely to fit a natural language query to the language of the system.

NELS IMPLEMENTATION ON A WIDE AREA NETWORK

As noted in Figure 1 above, NELS is built to operate over ethernet networks, running Network File Server (NFS) and Transaction Communication Protocol Internet Protocol (TCP/IP). This permits operation of NELS to be distributed on a local, a regional, or a national basis. Moreover, although NELS itself must run on a UNIX based operating system platform, files may be referenced and accessed anywhere that an eithernet connection is maintained. For example, it would be possible to have a system hosted on an IBM RISC 6000 that loaded an Interleaf viewer from a SUN Sparc and a document file from a DEC VAX. Both tools and files may be distributed on heterogeneous devices and drawn to a common platform for display. In this manner, data from many systems may be used from a single NELS host, with response times and access slots limited only by the power of the host device.

All data on all referenced systems is recorded in NELS through a "metadata" record. A metadata record is information about a file of documents, images, graphics, or drawings. The specific fields of the metadata may be customized by a librarian for specific collections. One of the fields in all metadata records is the "path" field which describes for NELS the directory locations of files on all other devices at all other locations. A typical metadata record for an image of the earth from space appears below as Figure 3.

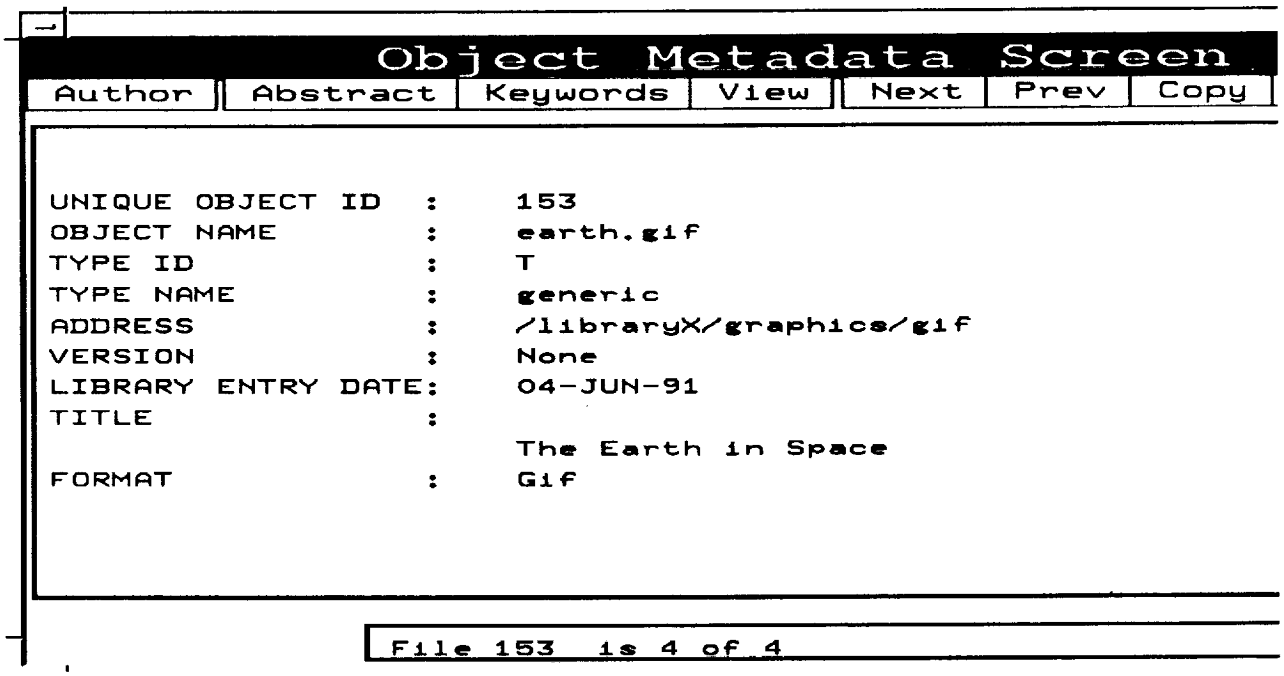


Figure 3: Metadata screen for a NELS entry.

Additonally, a number of devices hosting NELS may work in concert to optimize network topographies. For example, one network node may specialize in large numeric data files, another in image analyses, and another in engineering drawings. Although each node would be the primary user of its own data, however, by maintaining a complete metadata set at all three nodes, all data of all types could be accessed from any node, although more slowly at more remote locations. The diagram of Figure 4 below illustrates the distributed NELS concept.

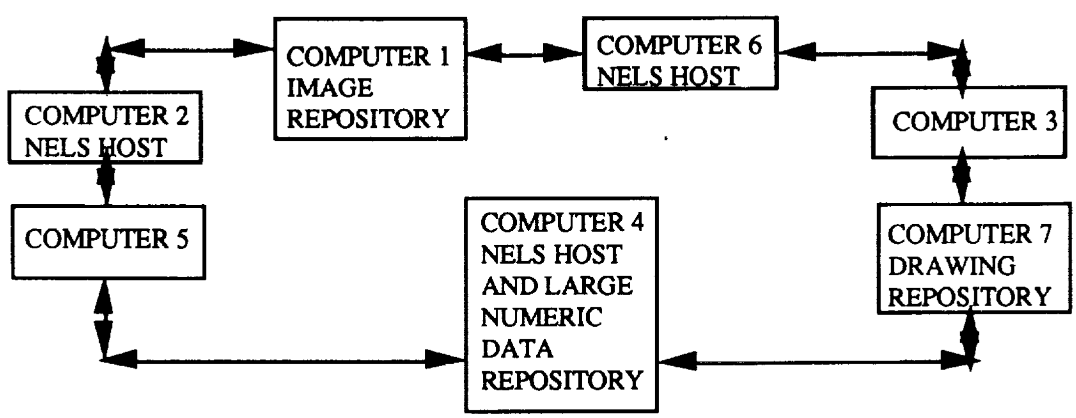


Figure 4: NELS distributed data base concept over a wide area network of computers.

NELS SCREEN EXAMPLES

NELS employs a Graphical User Interface (GUI) for screen display. With this type of interface, a user need only point with a mouse and click to select menu hierarchies, or operate screen buttons. Further, Motif window management software allows windows to be resized, or data to be viewed through the use of scroll bars. Finally, X-window/Motif software permits the inclusion of a number of file viewers. A viewer in this sense is a routine called by the window manager and passed the requested file as a parameter. Because of this, viewers may be added easily without disturbing the central application. Viewers presently supported in NELS 1.0 are ASCII, TIFF, GIF, raster, VE (a viewer for display of multiple images or bitmapped pages), and Interleaf. Viewers planned for implementation are autoCad, DECImage, and Post Script.

Figure 5 below displays the initial screen of NELS, which presents a list of the top hierarchies of the various sample libraries. A mouse click on any line would transfer the user to the next level of of organizational units, projects and offices. For known item searches, simply moving to the bottom of the list and displaying whatever objects are found there may remain the most efficient method of retrieval. The Function menu provides options for organizing the hierarchies by object class or aphabetically. Searching methods provided include natural language, query by example, and boolean logic. Collection defines searching depth, or the degree to which related collections are to be included in a given search. Admin lists functions available to Librarians such as adding and deleting collections and records, establishing users and user privileges and other functions. The screen overlayed in Figure 5 is the NELS search screen for natural language. The search request may be entered in simple English.

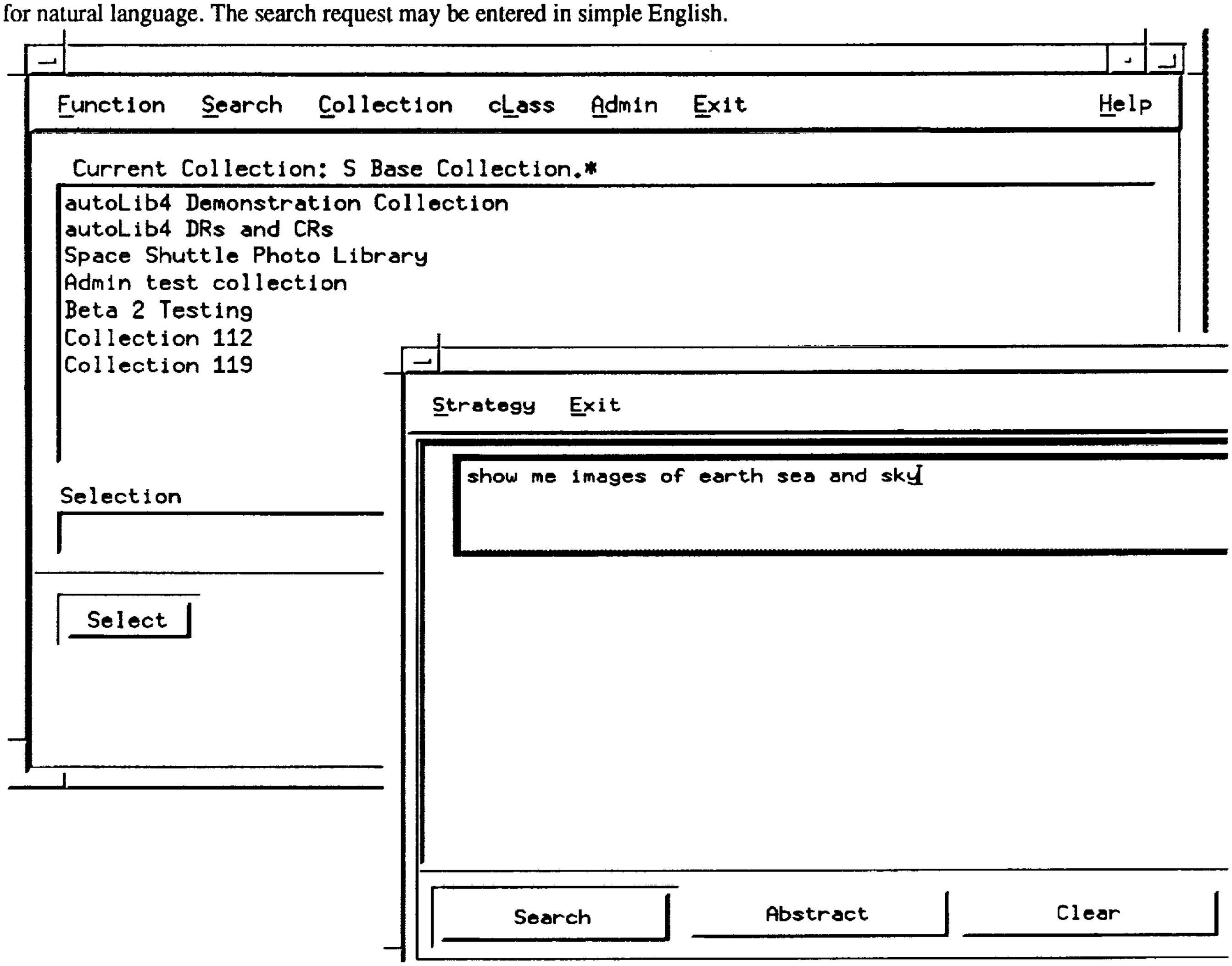


Figure 5: NELS main sign-on screen and natural language search screen.

Figure 6 below displays the ranked list of items resulting from the search for images made in Figure 5, with the retrieved image of the earth displayed in the lower left image corner. The numeric value shown to the left of the title on the object browser screen is the coefficient of correlation defined in the vector cosine formula shown earlier.

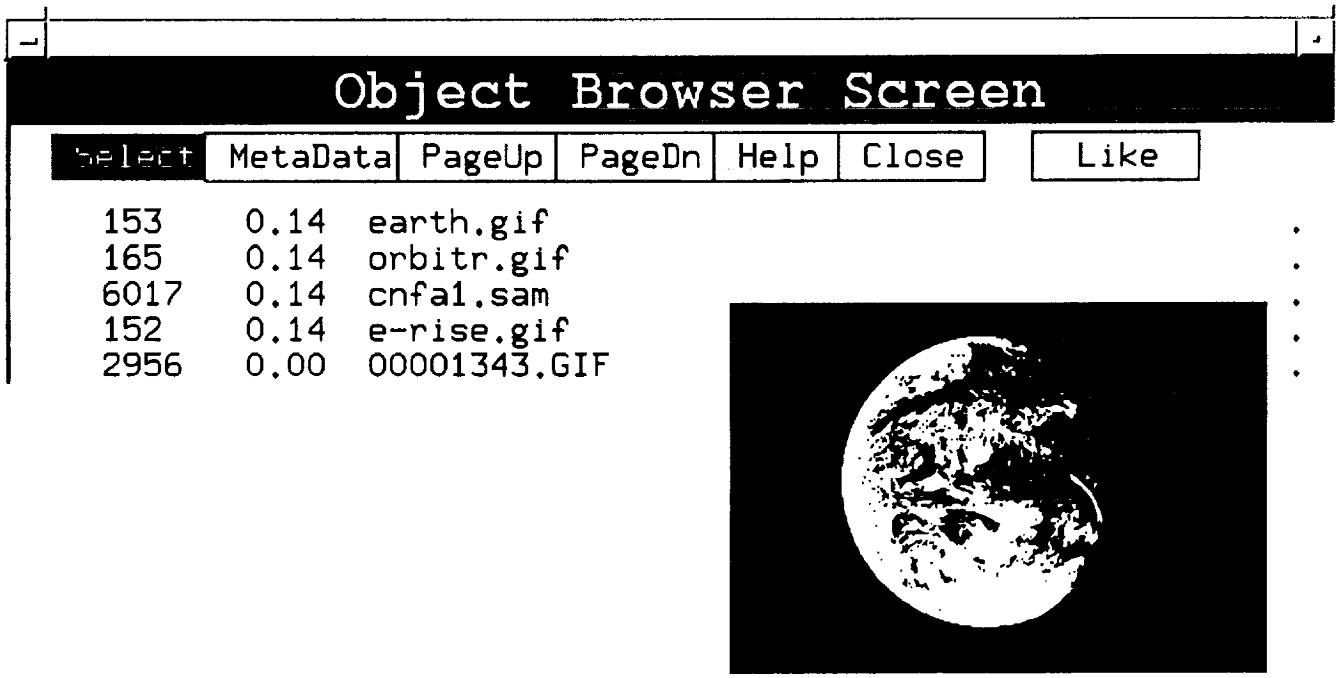


Figure 6: NELS object browser screen and retrieved image.

NELS SUPPORTED PLATFORMS

NELS is presently supported on the IBM RISC 6000 running under AIX, the DEC 5500 running under Ultrix, and the Sun 470 running under UNIX. Hewlett-Packard and Data General platforms are planned for future support. System access is supported for both general PC's and for Apple Mac II devices. Device classes for PC's are recommended to begin at the 386 level at 20 MHZ. Additionally, a minimum of 6MB of RAM memory is recommended. MAC II products are recommended to run System 6.05 minimally, with a minimum of 8MB RAM for best results. Both classes of devices require ethernet boards, however, these may obtained from a number of sources. X-software for PC's and Macintoshes has proliferated recently, and the market now offers a wide range of choices for this software as well [4].

REFERENCES

1. CLEVERDON, CYRIL W.; MILLS, J.; KEEN, MICHAEL. 1966. Factors Determining the Performance of Indexing Systems, Volume 1. Design. Volume 2. Test Results. Cranfield, England: ASLIB.

2. SALTON, GERARD; MCGILL,MICHAEL J. 1983. Introduction to Modern Information Retrieval. McGraw-Hill, New York.

3. RORVIG, MARK E. 1988 "Psychometric Measurement and Information Retrieval," In Annual Review of Information Science and Technology (ARIST), 23:157-189.

4. MCCOY, DANIEL J., 1991 "X Servers for PCs & Macintoshes," LibraryX, SoftwareTechnology Branch, NASA/Johnson Space Center, Loral Space Information Systems. NASA-JSC-PT4, Houston, Texas 77058. 1991

N92-22474

AUTOCLASS: AN AUTOMATIC CLASSIFICATION SYSTEM

John Stutz
NASA Ames Research Center
Artificial Intelligence Research Branch
Mail Stop 244-17
Moffett Field, CA 94035

Peter Cheeseman
Research Institute for Advanced Computer Science

Robin Hanson
Sterling Software

Abstract

The task of inferring a set of classes and class descriptions most likely to explain a given data set can be placed on a firm theoretical foundation using Bayesian statistics. Within this framework, and using various mathematical and algorithmic approximations, the AutoClass system searches for the most probable classifications, automatically choosing the number of classes and complexity of class descriptions. A simpler version of AutoClass has been applied to many large real data sets, have discovered new independently-verified phenomena, and have been released as a robust software package. Recent extensions allow attributes to be selectively correlated within particular classes, and allow classes to inherit, or share, model parameters though a class hierarchy. In this paper we summarize the mathematical foundations of Autoclass.

1 Introduction

The task of *supervised* classification - i.e., learning to predict class memberships of test cases given labeled training cases - is a familiar machine learning problem. A related problem is *unsupervised* classification, where training cases are also unlabeled. Here one tries to predict all features of new cases; the best classification is the least "surprised" by new cases. This type of classification, related to clustering, is often very useful in exploratory data analysis, where one has few preconceptions about what structures new data may hold.

We have previously developed and reported on AutoClass [Cheeseman *et al.*, 1988a; Cheeseman *et al.*, 1988b], an unsupervised classification system based on Bayesian theory. Rather than just partitioning cases, as most clustering techniques do, the Bayesian approach searches in a model space for the "best" class descriptions. A best classification optimally trades off predictive accuracy against the complexity of the classes, and so does not "overfit" the data. Such classes are also "fuzzy"; instead of each case being assigned to a class, a case has a probability of being a member of each of the different classes.

*Research Institute for Advanced Computer Science

Autoclass III, the most recent released version, combines real and discrete data, allows some data to be missing, and automatically chooses the number of classes from first principles. Extensive testing has indicated that it generally produces significant and useful results, but is primarily limited by the simplicity of the models it uses, rather than, for example, inadequate search heuristics. AutoClass III assumes that all attributes are relevant, that they are independent of each other within each class, and that classes are mutually exclusive. Recent extensions, embodied in Autoclass IV, let us relax two of these assumptions, allowing attributes to be selectively correlated and to have more or less relevance via a class hierarchy.

This paper summarizes the mathematical foundations of AutoClass, beginning with the Bayesian theory of learning, and then applying it to increasingly complex classification problems, from various single class models up to hierarchical class mixtures. For each problem, we describe our assumptions in words and mathematics, and then give the resulting evaluation and estimation functions for comparing models and making predictions. The derivations of these results from these assumptions, however, are not given.

2 Bayesian Learning

Bayesian theory gives a mathematical calculus of degrees of belief, describing what it means for beliefs to be consistent and how they should change with evidence. This section briefly reviews that theory, describes an approach to making it tractable, and comments on the resulting tradeoffs. In general, a Bayesian agent uses a single real number to describe its degree of belief in each proposition of interest. This assumption, together with some other assumptions about how evidence should affect beliefs, leads to the standard probability axioms. This result was originally proved by Cox [Cox, 1946] and has been reformulated for an AI audience [Heckerman, 1990]. We now describe this theory.

2.1 Theory

Let E denote some evidence that is known or could potentially be known to an agent; let H denote a hypothesis specifying that the world is in some particular state; and let the sets of possible evidence E and possible states of the world H each be mutually exclusive and exhaustive sets. For example, if we had a coin that might be two-headed the possible states of the world might be "ordinary coin", "two-headed coin". If we were to toss it once the possible evidence would be "lands heads", "lands tails".

In general, $P(ab|cd)$ denotes a real number describing an agent's degree of belief in the conjunction of propositions a and b, conditional on the assumption that propositions c and d are true. The propositions on either side of the conditioning bar "|" can be arbitrary Boolean expressions. More specifically, $\pi(H)$ is a "prior" describing the agent's belief in H *before*, or in the absence of, seeing evidence E, $\pi(H|E)$ is a "posterior" describing the agent's belief *after* observing some particular evidence E, and $L(E|H)$ is a "likelihood" embodying the agent's theory of how likely it would be to see each possible evidence combination E in each possible world H.

To be consistent, beliefs must be non-negative, $0 \leq P(a|b) \leq 1$, and normalised, so that $\sum_H \pi(H) = 1$ and $\sum_E L(E|H) = 1$. That is, the agent is sure that the world is in *some* state and that some evidence will be observed. The likelihood and the prior together give a "joint" probability $J(EH) \equiv L(E|H)\pi(H)$ of both E and H. Normalising the joint gives Bayes' rule, which tells how beliefs should change with evidence;

$$\pi(H|E) = \frac{J(EH)}{\sum_H J(EH)} = \frac{L(E|H)\pi(H)}{\sum_H L(E|H)\pi(H)}.$$

When the set of possible Hs is continuous, the prior $\pi(H)$ becomes a differential $d\pi(H)$, and the sums over H are replaced by integrals. Similarly, continuous Es have a differential likelihood $dL(E|H)$, though any real evidence ΔE will have a finite probability $\Delta L(E|H) \approx dL(E|H)\frac{\Delta E}{dE}$.

In theory, all an agent needs to do in any given situation is to choose a set of states H, an associated likelihood function describing what evidence is expected to be observed in those states, a set of prior expectations on the states, and then collect some relevant evidence. Bayes' rule then specifies the appropriate posterior beliefs about the state of the world, which can be used to answer most questions of interest. An agent can combine these posterior beliefs with its utility over states $U(H)$, which says how much it prefers each possible state, to choose an action A which maximises its expected utility

$$EU(A) = \sum_H U(H)\pi(H|EA).$$

2.2 Practice

In practice this theory can be difficult to apply, as the sums and integrals involved are often mathematically intractable. So one must use approximations. Here is our approach.

Rather than consider all possible *states* of the world, we focus on some smaller space of *models*, and do all of our analysis conditional on an assumption S that the world really is described by one of the models in our space. As with most modeling, this assumption is almost certainly false, but it makes the analysis tractable. With time and effort we can make our models more complex, expanding our model space in order to reduce the effect of this simplification.

The parameters which specify a particular model are split into two sets. First, a set of discrete parameters T describe the general form of the model, usually by specifying some functional form for the likelihood function. For example, T might specify whether two variables are correlated or not, or how many classes are present in a classification. Second, free variables in this general form, such as the magnitude of the correlation or the relative sizes of the classes, constitute the remaining continuous model parameters V.

We generally prefer a likelihood[1] $L(E|VTS)$ which is mathematically simple and yet still embodies the kinds of complexity we believe to be relevant.

Similarly, we prefer a simple prior distribution $d\pi(VT|S)$ over this model space, allowing the resulting V integrals, described below, to be at least approximated. A prior that predicts the different parameters in V independently, through a product of terms for each different parameter, often helps. We also prefer the prior to be as broad and uninformative as possible, so our software can be used in many different problem contexts, though in principal we could add specific domain knowledge through an appropriate prior. Finally we prefer a prior that gives nearly equal weight to different levels of model complexity, resulting in a "significance test". Adding more parameters to a model then induces a cost, which must be paid for by a significantly better fit to the data before the more complex model is preferred.

Sometimes the integrable priors are not broad enough, containing meta-parameters which specify some part of model space to focus on, even though we have no prior expectations about where to focus. In these cases we "cheat" and use simple statistics collected from the evidence we are going to use, to help set these priors[2]. For example, see Sections 4.2, 4.5.

The joint can now be written as $dJ(EVT|S) = L(E|VTS)\,d\pi(VT|S)$ and, for a reasonably-complex problem, is usually a very rugged distribution in VT, with an immense number of sharp peaks distributed widely over a huge high-dimensional space. Because of this we despair of directly normalising the joint, as required by Bayes' rule, or of communicating the detailed shape of the posterior distribution.

Instead we break the continuous V space into regions R surrounding each sharp peak, and search until we tire for combinations RT for which the "marginal" joint

$$M(ERT|S) \equiv \int_{V \in R} dJ(EVT|S)$$

is as large as possible. The best few such "models" RT are then reported, even though it is usually almost certain that more probable models remain to be found.

Each model RT is reported by describing its marginal joint $M(ERT|S)$, its discrete parameters T, and estimates of typical values of V in the region R, like the mean estimate of V:

$$\mathcal{E}(V|ERTS) \equiv \frac{\int_{V \in R} V\, dJ(EVT|S)}{M(ERT|S)}$$

[1] Note that when a variable like V sits in a probability expression where a proposition should be, it stands for a proposition that the variable has a particular value.

[2] This is cheating because the prior is supposed to be independent of evidence.

or the V for which $dJ(EVT|S)$ is maximum in R. While these estimates are not invariant under reparameterizations of the V space, and hence depend on the syntax with which the likelihood was expressed, the peak is usually sharp enough that such differences don't matter.

Reporting only the best few models is usually justified, since the models weaker than this are usually many orders of magnitude less probable than the best one. The main reason for reporting models other than the best is to show the range of variation in the models, so that one can judge how different the better, not yet found, models might be.

The decision to stop searching for better models RT than the current best can often be made in a principled way by using estimates of how much longer it would take to find a better model, and how much better than model would be. If the fact that a data value is unknown might be informative, one can model "unknown" as just another possible (discrete) data value; otherwise the likelihood for an unknown value is just a sum over the possible known values.

To make predictions with these resulting models, a reasonable approximation is to average the answer from the best few peaks, weighted by the relative marginal joints. Almost all of the weight is usually in the best few, justifying the neglect of the rest.

2.3 Tradeoffs

Bayesian theory offers the advantages of being theoretically well-founded and empirically well-tested [Berger, 1985]. It offers a clear procedure whereby one can almost "turn the crank", modulo doing integrals and search, to deal with any new problem. The machinery automatically trades off the complexity of a model against its fit to the evidence. Background knowledge can be included in the input, and the output is a flexible mixture of several different "answers," with a clear and well-founded decision theory [Berger, 1985] to help one use that output.

Disadvantages include being forced to be explicit about the space of models one is searching in, though this can be good discipline. One must deal with some difficult integrals and sums, although there is a huge literature to help one here. And one must often search large spaces, though most any technique will have to do this and the joint probability provides a good local evaluation function. Finally, it is not clear how one can take the computational cost of doing a Bayesian analysis into account without a crippling infinite regress.

Some often perceived disadvantages of Bayesian analysis are really not problems in practice. Any ambiguities in choosing a prior are generally not serious, since the various possible convenient priors usually do not disagree strongly within the regions of interest. Bayesian analysis is not limited to what is traditionally considered "statistical" data, but can be applied to any space of models about how the world might be. For a general discussion of these issues, see [Cheeseman, 1990].

We will now illustrate this general approach by applying it to the problem of unsupervised classification.

3 Model Spaces Overview

3.1 Conceptual Overview

In this paper we deal only with attribute-value, not relational, data.[3] For example, medical cases might be described by medical forms with a standard set of entries or slots. Each slot could be filled only by elements of some known set of simple values, like numbers, colors, or blood-types. (In this paper, we will only deal with real and discrete attributes.)

We would like to explain this data as consisting of a number of classes, each of which corresponds to a differing underlying cause for the symptoms described on the form. For example, different patients might fall into classes corresponding to the different diseases they suffer from.

To do a Bayesian analysis of this, we need to make this vague notion more precise, choosing specific mathematical formulas which say how likely any particular combination of evidence would be. A natural way to do this is to say that there are a certain number of classes, that a random patient has a certain probability to come from each of them, and that the patients are distributed independently – once we know all about the underlying classes then learning about one patient doesn't help us learn what any other patient will be like.

In addition, we need to describe how each class is distributed. We need a "single class" model saying how likely any given evidence is, given that we know what class the patient comes from. Thus we build the multiclass model space from some other pre-existing model space, which can be arbitrarily complex. (In fact, much of this paper will be spend describing various single class models.) In general, the more complex each class can be, the less of a need there is to invoke multiple classes to explain the variation in the data.

The simplest way to build a single-class model is to predict each attribute independently, i.e., build it from attribute-specific models. A class has a distribution for each attribute and, if you know the class of a case, learning the values of one attribute doesn't help you predict the value of any other attributes. For real attributes one can use a standard normal distribution, characterized by some specific mean and a variance around that mean. For discrete attributes one can use the standard multinomial distribution, characterized by a specific probability for each possible discrete value.

Up to this point we have described the model space of Autoclass III. Autoclass IV goes beyond this by introducing correlation and inheritance. Correlation is introduced by removing the assumption that attributes are independent within each class. The simplest way to do this is to let all real attributes covary, and let all discrete attributes covary. The standard way for real attributes to covary is the multivariate normal, which basically says that there is some other set of attributes one could define, as linear combinations of the attributes given, which vary independently according to normal distributions. A simple way to let discrete attributes covary is to define one super-attribute whose possible values are all possible

[3] Nothing in principle prevents a Bayesian analysis of more complex model spaces that predict relational data.

combinations of the values of the attributes given.

If there are many attributes, the above ways to add correlation introduce a great many parameters in the models, making them very complex and, under the usual priors, much less preferable than simpler independent models. What we really want are simpler models which only allow partial covariance. About the simplest way to do this is to say that, for a given class, the attributes clump together in blocks of inter-related attributes. All the attributes in a block covary with each other, but not with the attributes in other blocks. Thus we can build a block model space from the covariant model spaces.

Even this simpler form of covariance introduces more parameters that the independent case, and when each class must have its own set of parameters, multiple classes are penalized more strongly. Attributes which are irrelevant to the whole classification, like a medical patient's favorite color, can be particularly costly. To reduce this cost, one can allow classes to share the specification of parameters associated with some of their independent blocks. Irrelevant attributes can then be shared by all classes at a minimum cost.

Rather than allow arbitrary combinations of classes to share blocks, it is simpler to organize the classes as leaves of a tree. Each block can be placed at some node in this tree, to be shared by all the leaves below that node. In this way different attributes can be explained at different levels of an abstraction hierarchy. For medical patients the tree might have "viral infections" near the root, predicting fevers, and some more specific viral disease near the leaves, predicting more disease specific symptoms. Irrelevant attributes like favorite-color would go at the root.

3.2 Notation Summary

For all the models to be considered in this paper, the evidence E will consist of a set of I cases, an associated set $\mathcal{K}$ of attributes, of size[4] K, and case attribute values X_{ik}, which can include "unknown." For example, medical case number 8, described as (age = 23, blood-type = $A, \ldots$), would have $X_{8,1} = 23, X_{8,2} = A$, etc.

In the next two sections we will describe applications of Bayesian learning theory to various kinds of models which could explain this evidence, beginning with simple model spaces and building more complex spaces from them. We begin with a single class. First, a single attribute is considered, then multiple independent attributes, then fully covariant attributes, and finally selective covariance. In the next section we combine these single classes into class mixtures. Table 1 gives an overview of the various spaces.

For each space S we will describe the continuous parameters V, any discrete model parameters T, normalized likelihoods $dL(E|VTS)$, and priors $d\pi(VT|S)$. Most spaces have no discrete parameters T, and only one region R, allowing us to usually ignore these parameters. Approximations to the resulting marginals $M(ERT|S)$ and estimates $\mathcal{E}(V|ERTS)$ will be given, but not derived. These will often be given in terms of general functions F, so that they may be reused later on. As appropriate, comments will be made about algorithms and computational complexity. All of the likelihood functions considered here assume the cases are independent, i.e.,

$$L(E|VTS) = \prod_i L(E_i|VTS)$$

so we need only give $L(E_i|VTS)$ for each space, where $E_i \equiv \{X_{i1}, X_{i2}, X_{i3}, \ldots, X_{iK}\}$.

[4]Note we use script letters like $\mathcal{K}$ for sets, and matching ordinary letters K to denote their size.

4 Single Class Models

4.1 Single Discrete Attribute - S_{D1}

A discrete attribute k allows only a finite number of possible values $l \in [1, 2, ..., L]$ for any X_i. "Unknown" is usually treated here as just another possible value. A set of independent coin tosses, for example, might have $L = 3$ with l_1 = heads, l_2 = tails, and l_3 = "unknown". We make the assumption S_{D1} that there is only one discrete attribute, and that the only parameters are the continuous parameters $V = q_1 \ldots q_L$ consisting of the likelihoods $L(X_i|VS_{D1}) = q_{(l=X_i)}$ for each possible value l. In the coin example, $q_1 = .7$ would say that the coin was so "unbalanced" that it has a 70 percent chance of coming up heads each time.

There are only $L - 1$ free parameters since normalization requires $\sum_l q_l = 1$. For this likelihood, all that matters from the data are the number of cases with each value[5] $I_l = \sum_i \delta_{X_i l}$. In the coin example, I_1 would be the number of heads. Such sums are called "sufficient statistics" since they summarize all the information relevant to a model.

We choose a prior

$$d\pi(V|S_{D1}) = dB(q_1 \ldots q_L|L) \equiv \frac{\Gamma(aL)}{\Gamma(a)^L} \prod_l q_l^{a-1} dq_l$$

which for $a > 0$ is a special case of a beta distribution [Berger, 1985] ($\Gamma(y)$ is the Gamma function [Spiegel, 1968]). This formula is parameterized by a, a "hyperparameter" which can be set to different values to specify different priors. Here we set $a = 1/L$. This simple problem has only one maximum, whose marginal is given by

$$M(E|S_{D1}) = F_1(I_1, \ldots, I_L, I, L) \equiv \frac{\Gamma(aL) \prod_l \Gamma(I_l + a)}{\Gamma(aL + I)\Gamma(a)^L}$$

We have abstracted the function F_1, so we can refer to it later. The prior above was chosen because it has a form similar to the likelihood (and is therefore a "conjugate" prior), and to make the following mean estimate of q_l particularly simple

$$\mathcal{E}(q_l|ES_{D1}) = F_2(I_l, I, L) \equiv \frac{I_l + a}{I + aL} = \frac{I_l + \frac{1}{L}}{I + 1}$$

for $a = 1/L$. F_2 is also abstracted out for use later. Note that while $F_2(I_l, I, L)$ is very similar to the classical estimate of $\frac{I_l}{I}$, F_2 is defined even when $I = 0$. Using a hash table, these results can be computed in order I numerical steps, independent of L.

[5]Note that δ_{uv} denotes 1 when $u = v$ and 0 otherwise.

Space	*Description*	V	T	R	*Subspaces*	*Compute Time*
S_{D1}	Single Discrete	q_l				I
S_{R1}	Single Real	$\mu\sigma$				I
S_I	Independent Attrs	V_k			$S_1 \equiv S_{D1}$ or S_{R1}	IK
S_D	Covariant Discrete	$q_{l_1 l_2 \ldots}$				IK
S_R	Covariant Real	$\mu_k \Sigma_{kk'}$				$(I+K)K^2$
S_V	Block Covariance	V_b	$B\mathcal{K}_b$		$S_B \equiv S_D$ or S_R	$NK(I\overline{K_b}+K_b^2)$
S_M	Flat Class Mixture	$\alpha_c V_c$	C	R	$S_C \equiv S_I$ or S_V	$NK\overline{C}(I\overline{K_b}+K_b^2)$
S_H	Tree Class Mixture	$\alpha_c V_c$	$J_c\mathcal{K}_c T_c$	R	$S_C \equiv S_I$ or S_V	$NK\overline{C}(I\overline{K_b}+K_b^2)$

Table 1: Model Spaces

4.2 Single Real Attribute - S_{R1}

Real attribute values X_i specify a small range of the real line, with a center x_i and a precision, Δx_i, assumed to be much smaller than other scales of interest. For example, someone's weight might be measured as 70 ± 1 kilograms. For scalar attributes, which can only be positive, like weight, it is best to use the logarithm of that variable [Aitchison and Brown, 1957].

For S_{R1}, where there is only one real attribute, we assume the standard normal distribution, where the sufficient statistics are the data mean $\overline{x} = \frac{1}{I}\sum_i^I x_i$, the geometric mean precision $\widehat{\Delta x} = (\prod_i^I \Delta x_i)^{\frac{1}{I}}$ and the standard deviation s given by $s^2 = \frac{1}{I}\sum_i (x_i - \overline{x})^2$. V consists of a model mean μ and deviation σ, and the likelihood is given by the standard normal distribution.

$$dL(x_i|VS_{R1}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x_i - \mu}{\sigma})^2}\, dx_i.$$

For example, people's weight might be distributed with a mean of 80 kilograms and a deviation of 15. Since all real data have a finite width, we replace dx with Δx to approximate the likelihood $\Delta L(X_i|VS_{R1}) = \int_{\Delta x} dL(x_i|VS_{R1}) \cong \frac{\Delta x}{dx} dL(x_i|VS_{R1})$.

As usual, we choose priors that treat the parameters in V independently.

$$d\pi(V|S_{R1}) = d\pi(\mu|S_{R1})\, d\pi(\sigma|S_{R1})$$

We choose a prior on the mean to be flat in the range of the data,

$$d\pi(\mu|S_{R1}) = dR(\mu|\mu^+, \mu^-)$$

where $\mu^+ = \max x_i$, $\mu^- = \min x_i$, by using the general uniform distribution

$$dR(y|y^+, y^-) \equiv \frac{dy}{y^+ - y^-} \text{ for } y \in [y^-, y^+].$$

A flat prior is preferable because it is non-informative, but note that in order to make it normalizable we must cheat and use information from the data to cut it off at some point. In the single attribute case, we can similarly choose a flat prior in $\log(\sigma)$.

$$d\pi(\sigma|S_{R1}) = dR(\log(\sigma)|log(\Delta\mu), log(\min \Delta x_i))$$

where $\Delta\mu = \mu^+ - \mu^-$. The posterior again has just one peak, so there is only one region R, and the resulting marginal is

$$M(E|S_{R1}) = \frac{\sqrt{\pi}\,\Gamma(\frac{I-1}{2})}{2} \frac{1}{(\pi I)^{\frac{I}{2}}} \frac{1}{log(\Delta\mu/\min \Delta x_i)} \frac{\widehat{\Delta x}^I}{s^{I-1}\Delta\mu}.$$

Note that this joint is dimensionless. The estimates are simply $\mathcal{E}(\mu|ES_{R1}) = \overline{x}$, and $\mathcal{E}(\sigma|E) = \sqrt{\frac{I}{I+1}}s$. Computation here takes order I steps, used to compute the sufficient statistics.

4.3 Independent Attributes - S_I

We now introduce some notation for collecting sets of indexed terms like X_{ik}. A single such term inside a $\{\}$ will denote the set of all such indexed terms collected across all of the indices, like i and k in $E = \{X_{ik}\} \equiv \{X_{ik} \text{ such that } i \in [1, \ldots, I], k \in \mathcal{K}\}$. To collect across only some of the indices we use $\bigcup_k$ as in $E_i = \bigcup_k X_{ik} \equiv \{X_{i1}, X_{i2}, \ldots\}$, all the evidence for a single case i.

The simplest way to deal with cases having multiple attributes is to assume S_I that they are all independent, i.e., treating each attribute as if it were a separate problem. In this case, the parameter set V partitions into parameter sets $V_k = \bigcup_{l_k} q_{l_k}$ or $[\mu_k, \sigma_k]$, depending on whether that k is discrete or real. The likelihood, prior, and joint for multiple attributes are all simple products of the results above for one attribute: $S_1 = S_{D1}$ or S_{R1} — i.e.,

$$L(E_i|VS_I) = \prod_k L(X_{ik}|V_k S_1),$$

$$d\pi(V|S_I) = \prod_k d\pi(V_k|S_1),$$

and

$$M(E|S_I) = \prod_k J(E(k)|S_1)$$

where $E(k) \equiv \bigcup_i X_{ik}$, all the evidence associated with attribute k. The estimates $\mathcal{E}(V_k|ES_I) = \mathcal{E}(V_k|E(k)S_1)$ are exactly the same. Computation takes order IK steps here.

4.4 Fully Covariant Discretes - S_D

A model space S_D which allows a set $\mathcal{K}$ of *discrete* attributes to fully covary (i.e, contribute to a likelihood in non-trivial combinations) can be obtained by treating all combinations of base attribute values as particular values of one super attribute, which then has $L' = \prod_k L_k$ values — so L' can be a very large number! V consists of terms like $q_{l_1 l_2 \ldots l_K}$, indexed by all the attributes. I_l generalises to

$$I_{l_1 l_2 \ldots l_K} = \sum_i \prod_k \delta_{x_{ik} l_k}.$$

Given this transformation, the likelihoods, etc. look the same as before:

$$L(E_i \mid V S_D) = q_{l_1 l_2 \ldots l_K},$$

where each $l_k = X_{ik}$,

$$d\pi(V|S_D) = dB(\{q_{l_1 l_2 \ldots l_K}\} \mid L'),$$

$$M(E|S_D) = F_1(\{I_{l_1 l_2 \ldots l_K}\}, I, L'),$$

and [6]

$$\mathcal{E}(q_{l_1 l_2 \ldots l_K}|ES_D) = F_2(I_{l_1 l_2 \ldots l_K}, I, L')$$

Computation takes order IK steps here. This model could, for example, use a single combined hair-color eye-color attribute to allow a correlation between people being blond and blue-eyed.

4.5 Fully Covariant Reals - S_R

If we assume S_R that a set $\mathcal{K}$ of real-valued attributes follow the multivariate normal distribution, we replace the σ_k^2 above with a model covariance matrix $\Sigma_{kk'}$ and s_k^2 with a data covariance matrix

$$\mathrm{S}_{kk'} = \frac{1}{I} \sum_i (x_{ik} - \bar{x}_k)(x_{ik'} - \bar{x}_{k'})$$

. The $\Sigma_{kk'}$ must be symmetric, with $\Sigma_{kk'} = \Sigma_{k'k}$, and "positive definite", satisfying $\sum_{kk'} y_k \Sigma_{kk'} y_{k'} > 0$ for any vector y_k. The likelihood for a set of attributes $\mathcal{K}$ is[7]

$$\begin{aligned} dL(E_i|VS_R) &= dN(E_i, \{\mu_k\}, \{\Sigma_{kk'}\}, \mathcal{K}) \\ &\equiv \frac{e^{-\frac{1}{2}\sum_{kk'}(x_k-\mu_k)\Sigma^{\text{inv}}_{kk'}(x_{k'}-\mu_{k'})}}{(2\pi)^{\frac{K}{2}} |\Sigma_{kk'}|^{\frac{1}{2}}} \prod_k dx_k \end{aligned}$$

is the multivariate normal in K dimensions.

As before, we choose a prior that takes the means to be independent of each other, and independent of the covariance

$$d\pi(V|S_R) = d\pi(\{\Sigma_{kk'}\}|S_R) \prod_k d\pi(\mu_k|S_{R1}),$$

so the estimates of the means remain the same, $E(\mu_k|ES_R) = \bar{x}_k$. We choose the prior on $\Sigma_{kk'}$ to use an inverse Wishart distribution [Mardia *et al.*, 1979]

$$d\pi(\{\Sigma_{kk'}\}|S_R) = d\mathcal{W}^{\text{inv}}_K(\{\Sigma_{kk'}\} \mid \{\mathrm{G}_{kk'}\}, h) \equiv$$

$$\frac{|\mathrm{G}_{kk'}|^{-\frac{h}{2}} |\Sigma_{kk'}|^{\frac{-h-K-1}{2}} e^{-\frac{1}{2}\sum^K_{kk'} \Sigma^{\text{inv}}_{kk'} \mathrm{G}^{\text{inv}}_{k'k}}}{2^{\frac{Kh}{2}} \pi^{\frac{K(K-1)}{4}} \prod^K_a \Gamma(\frac{h+1-a}{2})} \prod^K_{k \le k'} d\Sigma_{kk'}$$

which is normalized (integrates to 1) for $h \geq K$ and $\Sigma_{kk'}$ symmetric positive definite. This is a "conjugate" prior, meaning that it makes the resulting posterior $d\pi(\{\Sigma_{kk'}\}|ES_R)$ take the same mathematical form as the prior. This choice makes the resulting integrals manageable, but requires us to choose an h and all the components of $\mathrm{G}_{kk'}$. We choose $h = K$ to make the prior as broad as possible, and for $\mathrm{G}_{kk'}$ we "cheat" and choose $\mathrm{G}_{kk'} = \mathrm{S}_{kk}\delta_{kk'}$ in order to avoid overly distorting the resulting marginal

$$M(E|S_R) = \frac{\prod^K_a \frac{\Gamma(\frac{I+h-a}{2})}{\Gamma(\frac{1+h-a}{2})}}{I^{\frac{K}{2}} \pi^{\frac{K(I-1)}{2}}} \frac{|\mathrm{G}_{kk'}|^{\frac{h}{2}}}{|I\mathrm{S}_{kk'} + \mathrm{G}_{kk'}|^{\frac{I+h-1}{2}}} \prod_k^K \frac{\widehat{\Delta x_k}^I}{\Delta \mu_k}$$

and estimates

$$E(\Sigma_{kk'}|ES_R) = \frac{I\mathrm{S}_{kk'} + \mathrm{G}_{kk'}}{I + h - K - 2} = \frac{I + \delta_{kk'}}{I - 2}\mathrm{S}_{kk'}.$$

If we choose $\mathrm{G}_{kk'}$ too large it dominates the estimates, and if $\mathrm{G}_{kk'}$ is too small the marginal is too small. The compromise above should only over estimate the marginal somewhat, since it in effect pretends to have seen previous data which agrees with the data given. Note that the estimates are undefined unless $I > 2$. Computation here takes order $(I + K)K^2$ steps. At present, we lack a satisfactory way to approximate the above marginal when some values are unknown.

4.6 Block Covariance - S_V

Rather than just having either full independence or full dependence of attributes, we prefer a model space S_V where some combinations of attributes may covary while others remain independent. This allows us to avoid paying the cost of specifying covariance parameters when they cannot buy us a significantly better fit to the data.

We partition the attributes $\mathcal{K}$ into B blocks $\mathcal{K}_b$, with full covariance within each block and full independence between blocks. Since we presently lack a model allowing different types of attributes to covary, all the attributes in a block must be of the same type. Thus real and discretes may not mutually covary.

We are away of other models of partial dependence, such as the the trees of Chow and Liu described in [Pearl, 1988], but choose this approach because it includes the limiting cases of full dependence and full independence.

The evidence E partitions block-wise into $E(\mathcal{K}_b)$ (using $E_i(\mathcal{K}) \equiv \bigcup_{k \in \mathcal{K}} X_{ik}$ and $E(\mathcal{K}) \equiv \{E_i(\mathcal{K})\}$), each with its own sufficient statistics; and the parameters V partition into parameters $V_b = \{q_{l_1 l_2 \ldots l_K}\}$ or $[\{\Sigma_{kk'}\}, \{\mu_k\}]$. Each block is treated as a different problem, except that we now also have discrete parameters T to specify which attributes covary, by specifying B blocks and $\{\mathcal{K}_b\}$ attributes in each block. Thus the likelihood

$$L(E_i|VTS_V) = \prod_b^B L(E_i(\mathcal{K}_b)|V_b S_B)$$

is a simple product of block terms $S_B = S_D$ or S_R assuming full covariance within each block, and the estimates $\mathcal{E}(V_b|ETS_V) = \mathcal{E}(V_b|E(\mathcal{K}_b)S_B)$ are the same as before.

We choose a prior which predicts the block structure $B\{\mathcal{K}_b\}$ independently of the parameters V_b within each independent block

$$d\pi(VT|S_V) = \pi(B\{\mathcal{K}_b\}|S_V) \prod_b d\pi(V_b|S_B)$$

---

[6] F_1 and F_2 are defined on page 4.

[7] Σ^{inv}_{ab} denotes the matrix inverse of Σ_{ab} satisfying $\sum_b \Sigma^{\text{inv}}_{ab}\Sigma_{bc} = \delta_{ac}$, and $|\Sigma_{ab}|$ denotes components of the matrix determinant of $\{\Sigma_{ab}\}$.

which results in a similarly decomposed marginal

$$M(ET|S_V) = \pi(B\,\{\mathcal{K}_b\}\,|S_V) \prod_b M(E(\mathcal{K}_b)|S_B).$$

We choose a block structure prior

$$\pi(B\,\{\mathcal{K}_b\}\,|S_V) = 1/K_R Z(K_R, B_R) K_D Z(K_D, B_D),$$

where $\mathcal{K}_R$ is the set of real attributes and B_R is the number of real blocks (and similarly for $\mathcal{K}_D$ and B_D). This says that it is equally likely that there will be one or two or three, etc. blocks, and, given the number of blocks, each possible way to group attributes is equally likely. This is normalized using $Z(A,U)$, given by

$$Z(A,U) \equiv \sum_{u=1}^{U} (-1)^{u-1} \frac{(U-u+1)^A}{(U-u+1)!\,(u-1)!},$$

which gives the number of ways one can partition a set with A elements into U subsets. This prior prefers the special cases of full covariance and full independence, since there are fewer ways to make these block combinations. For example, in comparing the hypothesis that each attribute is in a separate block (i.e., all independent) with the hypothesis that only one particular pair of attributes covary together in a block of size two, this prior will penalize the covariance hypothesis in proportion to the number of such pairs possible. Thus this prior includes a "significance test", so that a covariance hypothesis will only be chosen if the added fit to the data from the extra covariance is enough to overcome this penalty.

Computation here takes order $NK(I\overline{K_b} + \overline{K_b^2})$ steps, where N is the number of search trials done before quitting, which would be around $(K-1)!$ for a complete search of the space. $\overline{K_b}$ is an average, over both the search trials and the attributes, of the block size of real attributes (and unity for discrete attributes).

5 Class Mixtures

5.1 Flat Mixtures - S_M

The above model spaces $S_C = S_V$ or S_I can be thought of as describing a single class, and so can be extended by considering a space S_M of simple mixtures of such classes [D.M.Titterington *et al.*, 1985]. Figure 1 shows how this model, with $S_C = S_I$, can fit a set of artificial real-valued data in five dimensions.

In this model space the likelihood

$$L(E_i|VTS_M) = \sum_c^C \alpha_c L(E_i|V_c T_c S_C)$$

sums over products of "class weights" α_c, that give the probability that any case would belong to class c of the C classes, and class likelihoods describing how members of each class are distributed. In the limit of large C this model space is general enough to be able to fit any distribution arbitrarily closely, and hence is "asymtotically correct".

The parameters $T = [C, \{T_c\}]$ and $V = [\{\alpha_c\}, \{V_c\}]$ combine parameters for each class and parameters describing the mixture. The prior is similarly broken down as

$$d\pi(VT|S_M) = F_3(C)C!\; dB(\{\alpha_c\}\,|C) \prod_c d\pi(V_c T_c|S_C)$$

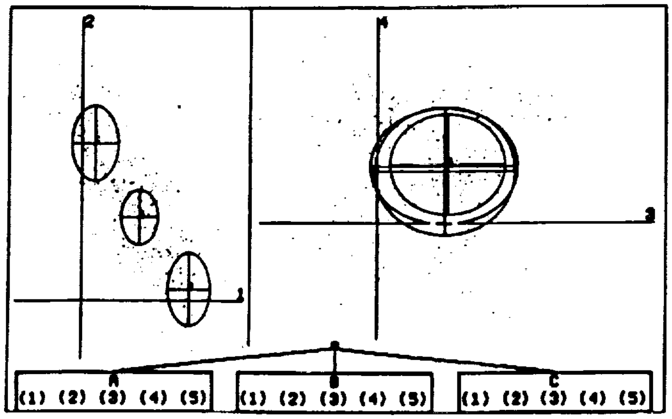


Figure 1: AutoClass III Finds Three Classes

We plot attributes 1 vs. 2, and 3 vs. 4 for an artificial data set. One σ deviation ovals are drawn around the centers of the three classes.

where $F_3(C) \equiv \frac{6}{\pi^2 C^2}$ for $C > 0$ and is just one arbitrary choice of a broad prior over integers. The α_c is treated as if the choice of class were another discrete attribute, except that a $C!$ is added because classes are not distinguishable a priori.

Except in very simple problems, the resulting joint $dJ(EVT|S)$ has many local maxima, and so we must now focus on regions R of the V space. To find such local maxima we use the "EM" algorithm [Dempster *et al.*, 1977] which is based on the fact that at a maxima the class parameters V_c can be estimated from weighted sufficient statistics. Relative likelihood weights

$$w_{ic} = \frac{\alpha_c L(E_i|V_c T_c S_C)}{L(E_i|VTS_M)},$$

give the probability that a particular case i is a member of class c. These weights satisfy $\sum_c w_{ic} = 1$, since every case must really belong to one of the classes. Using these weights we can break each case into "fractional cases", assign these to their respective classes, and create new "class data" $E^c = \bigcup_{ik} [X_{ik}, w_{ic}]$ with new weighted-class sufficient statistics obtained by using weighted sums $\sum_i w_{ic}$ instead of sums $\sum_i$. For example $I_c = \sum_i w_{ic}$, $\overline{x}_{kc} = \frac{1}{I_c} \sum_i w_{ic} x_{ik}$, $I_{l_1 \ldots l_K c} = \sum_i w_{ic} \prod_k \delta_{x_{ik} l_k}$, and $\widehat{\Delta x_{kc}} = \prod_i^I \Delta x_{ik}^{\frac{w_{ic}}{I_c}}$. Substituting these statistics into any previous class likelihood function $L(E|V_c T_c S_C)$ gives a weighted likelihood $L'(E^c|V_c T_c S_C)$ and associated new estimates and marginals.

At the maxima, the weights w_{ic} should be consistent with estimates of $V = \{[\alpha_c, C_c]\}$ from $\mathcal{E}(V_c|ERS_M) = \mathcal{E}'(V_c|E^c S_C)$ and $\mathcal{E}(\alpha_c|ERS_M) = F_2(I_c, I, C)$. To reach a maxima we start out at a random seed and repeatedly use our current best estimates of V to compute the w_{ic}, and then use the w_{ic} to re-estimate the V, stopping when they both predict each other. Typically this takes $10 - 100$ iterations. This procedure will converge from any starting point, but converges more slowly near the peak than second-order methods.

Integrating the joint in R can't be done directly because the product of a sum in the full likelihood is hard to decompose, but if we use fractional cases to approximate the likelihood

$$
\begin{aligned}
L(E_i|VTRS_m) &= \sum_c^C \alpha_c L(E_i|V_cT_cS_c) \\
&\cong \prod_c (\alpha_c L(E_i|V_cT_cS_C))^{w_{ic}}
\end{aligned}
$$

holding the w_{ic} fixed, we get an approximate joint:

$$M(ERT|S_M) \cong F_3(C)C!\,F_1(\{I_c\},I,C)\prod_c M'(E^cT|S_C)$$

Our standard search procedure combines an explicit search in C with a random search in all the other parameters. Each trial begins converging from classes built around C random case pairs. The C is chosen randomly from a log-normal distribution fit to the Cs of the $6-10$ best trials seen so far, after trying a fixed range of Cs to start. We also have developed alternative search procedures which selectively merge and split classes according to various heuristics. While these usually do better, they sometimes do much worse.

The marginal joints of the different trials generally follow a log-normal distribution, allowing us to estimate during the search how much longer it will take on average to find a better peak, and how much better it is likely to be.

In the simpler model space S_{MI} where $S_C = S_I$ the computation is order $NI\overline{C}K$, where $\overline{C}$ averages over the search trials. N is the number of possible peaks, out of the immense number usually present, that a computation actually examines. In the covariant space S_{MV} where $S_C = S_V$ this becomes $NK\overline{C}(I\overline{K_b} + \overline{K_b^2})$.

5.2 Class Hierarchy and Inheritance - S_H

The above class mixture model space S_M can be generalized to a hierarchical space S_H by replacing the above set of classes with a tree of classes. Leaves of the tree, corresponding to the previous classes, can now inherit specifications of class parameters from "higher" (closer to the root) classes. For the purposes of the parameters specified at a class, all of the classes below that class pool their weight into one big class. Parameters associated with "irrelevant" attributes are specified independently at the root. Figure 2 shows how a class tree, this time with $S_C = S_V$, can better fit the same data as in Figure 1. See [Hanson *et al.*, 1991] for more about this comparison.

The tree of classes has one root class r. Every other class c has one parent class P_c, and every class has J_c child classes given by C_{cj}, where the index j ranges over the children of a class. Each child class has a weight α_{cj} relative to its siblings, with $\sum_j^{J_c} \alpha_{cj} = 1$, and an absolute weight $\alpha_{C_{cj}} = \alpha_{cj}\alpha_c$, with $\alpha_r = 1$.

While other approaches to inheritance are possible, here each class is given an associated set of attributes $\mathcal{K}_c$, which it predicts independently through a likelihood $L(E_i(\mathcal{K}_c)|V_cT_cS_C)$ and which no class above or below it predicts. To avoid having redundant trees which describe the same likelihood function, only $\mathcal{K}_r$ can be empty, and non-leaves must have $J_c \geq 2$.

We need to ensure that all attributes are predicted somewhere at or above each leaf class. So we call $\mathcal{A}_c$

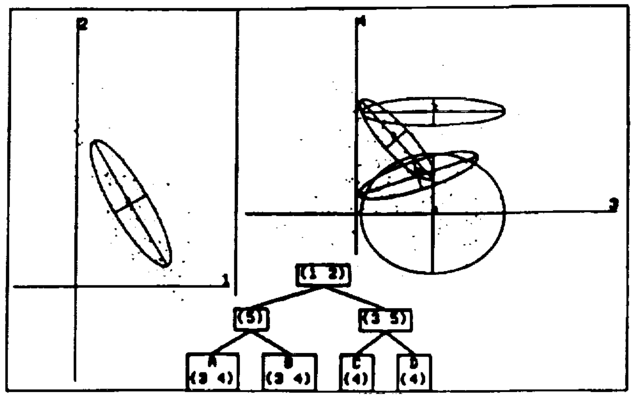


Figure 2: AutoClass IV Finds Class Tree $\times 10^{120}$ Better
Lists of attribute numbers denote covariant blocks within each class, and the ovals now indicate the leaf classes.

the set of attributes which are predicted at or below each class, start with $\mathcal{A}_r = K$, and then recursively partition each $\mathcal{A}_c$ into attributes $\mathcal{K}_c$ "kept" at that class, and hence predicted directly by it, and the remaining attributes to be predicted at or below each child $\mathcal{A}_{C_{cj}}$. For leaves $\mathcal{A}_c = \mathcal{K}_c$.

Expressed in terms of the leaves the likelihood is again a mixture:

$$L(E_i|VTS_M) = \sum_{c:J_c=0} \alpha_c \prod_{c'=c,P_c,P_{P_c},\ldots,r} L(E_i(\mathcal{K}_{c'})|V_{c'}T_{c'}S_C)$$

allowing the same EM procedure as before to find local maximas. The case weights here $w_{ci} = \sum_j^{J_c} w_{C_{cj}i}$ (with $w_{ri} = 1$) sum like in the flat mixture case and define class statistics $E^c(\mathcal{K}_c) = \bigcup_{k\in\mathcal{K}_c,i}[X_{ik}, w_{ci}]$.

We also choose a similar prior, though it must now specify the $\mathcal{K}_c$ as well:

$$d\pi(VT|S_H) =$$

$$\prod_c d\pi(J_c\mathcal{K}_c \mid \mathcal{A}_cS_H)J_c!\; dB(\bigcup_j \alpha_{cj}|J_c)\, d\pi(V_cT_c \mid \mathcal{K}_cS_C)$$

$$d\pi(J_c\mathcal{K}_c \mid \mathcal{A}_cS_H) = F_3(J_c-1)\frac{\mathcal{K}_c!\,(\mathcal{A}_c-\mathcal{K}_c)!}{(\mathcal{A}_c+\delta_{rc})\mathcal{A}_c!}$$

for all subsets $\mathcal{K}_c$ of $\mathcal{A}_c$ of size in the range $[1-\delta_{cr}, \mathcal{A}_c]$, except that $F_3(J_c-1)$ is replaced by δ_{0J_c} when $\mathcal{A}_c = \mathcal{K}_c$. Note that this prior is recursive, as the prior for each class depends on the what attributes have been chosen for its parent class.

This prior says that each possible number of attributes kept is equally likely, and given the number to be kept each particular combination is equally likely. This prior prefers the simpler cases of $\mathcal{K}_c = \mathcal{A}_c$ and $\mathcal{K}_c = 1$ and so again offers a significance test. In comparing the hypothesis that all attributes are kept at class with a hypothesis that all but one particular attribute will be kept at that class, this prior penalizes the all-but-one hypothesis in proportion to the number of attributes that could have been kept instead.

The marginal joint becomes

$$M(ERT|S_H) \cong$$

$$\prod_c d\pi(J_c\mathcal{K}_c \mid \mathcal{A}_cS_H)J_c!\,F_1(\bigcup_j I_{C_{cj}}, I_c, J_c)M'(E^c(\mathcal{K}_c)T_c|S_C)$$

and estimates are again $\mathcal{E}(V_c|ERS_H) = \mathcal{E}'(V_c|E^c(\mathcal{K}_c)S_C)$ and $\mathcal{E}(\alpha_{cj}|ERS_H) = F_2(I_{cj}, I_c, J_c)$.

In the general case of S_{HV}, where $S_C = S_V$, computation again takes $NK\overline{C}(I\overline{K_b} + \overline{K_b^2})$, except that the $\overline{J}$ is now also an average of, for each k, the number of classes in the hierarchy which use that k (i.e., have $k \in \mathcal{K}_c$). Since this is usually less than the number of leaves, the model S_H is typically cheaper to compute than S_M for the same number of leaves.

Searching in this most complex space S_{HV} is challenging. There are a great many search dimensions where one can trade off simplicity and fit to the data, and we have only begun to explore possible heuristics. Blocks can be merged or split, classes can be merged or split, blocks can be promoted or demoted in the class tree, EM iterations can be continued farther, and one can try a random restart to seek a new peak. But even the simplest approaches to searching a more general model space seem to do better than smarter searches of simpler spaces.

6 Conclusion

The Bayesian approach to unsupervised classification describes each class by a likelihood function with some free parameters, and then adds in a few more parameters to describe how those classes are combined. Prior expectations on those parameters VT combine with the evidence E to produce a marginal joint $M(ERT|S)$ which is used as an evaluation function for classifications in a region R near some local maxima of the continuous parameters V and with some choice of discrete model parameters T. This evaluation function optimally trades off the complexity of the model with its fit to the data, and is used to guide an open-ended search for the best classification.

In this paper we have applied this theory to model spaces of varying complexity in unsupervised classification. For each space we provides a likelihood, prior, marginal joint, and estimates. This should provide enough information to allow anyone to reproduce AutoClass, or to use the same evaluation functions in other contexts where these models might be relevant.

References

[Aitchison and Brown, 1957] J. Aitchison and J.A.C. Brown. *The Lognormal Distribution*. Cambridge at the University Press, 1957.

[Berger, 1985] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 1985.

[Cheeseman *et al.*, 1988a] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman. Autoclass: a Bayesian classification system. In *Proceedings of the Fifth International Conference on Machine Learning*, 1988.

[Cheeseman *et al.*, 1988b] P. Cheeseman, M. Self, J. Kelly, J. Stutz, W. Taylor, and D. Freeman. Bayesian classification. In *Seventh National Conference on Artificial Intelligence*, pages 607–611, Saint Paul, Minnesota, 1988.

[Cheeseman, 1990] Peter Cheeseman. On finding the most probable model. In Jeff Shrager and Pat Langley, editors, *Computational Models of Discovery and Theory Formation*, pages 73–96. Morgan Kaufmann, Palo Alto, 1990.

[Cox, 1946] R.T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1), 1946.

[Dempster *et al.*, 1977] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39:1–38, 1977.

[D.M.Titterington *et al.*, 1985] D.M.Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, New York, 1985.

[Hanson *et al.*, 1991] R. Hanson, J. Stutz, and P. Cheeseman. Bayesian classification with correlation and inheritance. In *12th International Joint Conference on Artificial Intelligence*, Sydney, Australia, 1991.

[Heckerman, 1990] David Heckerman. Probabilistic interpretations for mycin's certainty factors. In Glenn Shafer and Judea Pearl, editors, *Readings in Uncertain Reasoning*, pages 298–312. Morgan Kaufmann, San Mateo, California, 1990.

[Mardia *et al.*, 1979] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Academic Press Inc., New York, 1979.

[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgain Kaufmann, San Mateo, California, 1988.

[Spiegel, 1968] Murray Spiegel. *Mathematical Handbook of Formulas and Tables*. McGraw-Hill Book Company, New York, 1968.

N92-22475

SILVABASE: A FLEXIBLE DATA FILE MANAGEMENT SYSTEM

Steven J. Lambing
NASA Marshall Space Flight Center
Huntsville, AL 35812

Sandra J. Reynolds
Boeing Computer Support Services
Huntsville, AL 35812

ABSTRACT

The need for a more flexible and efficient data file management system for mission planning in the Mission Operations Laboratory (EO) at the Marshall Space Flight Center (MSFC) has spawned the development of Silvabase. Silvabase is a new data file structure based on a B+ tree data structure. This data organization allows for efficient forward and backward sequential reads, random searches and appends to existing data. It also provides random insertions and deletions with reasonable efficiency, utilization of storage space well but not at the expense of speed, and performance of these functions on a large volume of data. Mission planners required that some data be keyed and manipulated in ways not found in a commercial product. Mission planning software is currently being converted to use Silvabase in the Spacelab and Space Station Mission Planning Systems. Silvabase runs on a Digital Equipment Corporation's popular VAX/VMS computers in VAX FORTRAN. Silvabase has unique features involving time histories and intervals such as in operations research. Because of its flexibility and unique capabilities, Silvabase could be used in almost any government or commercial application that requires efficient reads, searches, and appends in medium to large amounts of almost any kinds of data.

INTRODUCTION

Context

In the Mission Operations Laboratory at the Marshall Space Flight Center, payload operations for Spacelab and non-Spacelab shuttle flights are planned and conducted. Mission planning includes projecting and designing the Space Shuttle Orbiter's trajectory and attitudes, determining from that the periods of opportunities for execution of payload experiments, scheduling those payload operations, and scheduling and managing the two way flow of scientific data between the orbiter and the ground. This endeavor calls for a variety of forms and amounts of data to be stored in secondary computer storage (disk files). Mission Planners require that their software be able to efficiently read this data sequentially forward and backward, efficiently search for a random key, and append new records to the existing sets of records. Additionally, they require the ability to make random insertions and deletions, and utilize storage space well without adversely impacting access speed.

The forms that all of this mission planning data takes are not homogeneous. Some of the data is simply a collection of related data items of different data types. Other data takes a tabular form consisting of numerous records, each having the same format (being homogeneous) with one data field serving as the key and usually representing time. There are numerous other forms. Additionally, mission planners have one unique type of data and methods of manipulating it. This data consists of a key which is made up of two values. The two values represent a starting time and an ending time of a time interval. Associated with this interval key, or these "On/Off" times, will be zero or more other data values, depending on what event the interval represents. A set of records, keyed by these intervals, are used to represent an intermittent recurring event or set of conditions. This type of data representation is frequently used in mission planning.

History

During the latter 1970s and the 1980s, mission planners stored much of their data in a set of file formats known as MIPS files. MIPS stands for the Marshall Interactive Planning System. This term is used to refer to all of, or several different components of, the Mission Operations Laboratory's mission planning computers and software. For the purposes of this paper however, MIPS only refers to these file formats and associated file access software used by mission planners before the advent of Silvabase. MIPS development began around 1974 on a Sperry 1100 computer. When Digital Equipment Corporation's VAX computers were obtained in 1979, MIPS was migrated over to that platform. In 1986, mission planners began to question the future of MIPS. It had become problematic because it was not originally designed for the VAX nor was it implemented in such a way as to be maintainable. Software analysts determined that minor fixes to MIPS problems were not cost effective due to decreasing confidence in it and major modifications would be equivalent to a rewrite [1]. MIPS had served the mission planners well, but it was time to move on.

It was determined to replace the MIPS file system with a new data file system that was designed to work in the VAX/VMS environment and address all of the mission planning requirements. An obvious option, the use of a commercial database system, was rejected. A commercial database could not be used because there were indications that data retrieval and record insertion would be too slow. Also, the capability to correctly handle interval-keyed data and provide the special functions for interval-keyed data required by mission planners was not found in any commercial product. Finally, and possibly most importantly, it must be possible to freely distribute mission planning software. Other organizations outside of MSFC's Mission Operations Laboratory have been and will in the future be required to use the mission planning software. If a commercial product of any kind was required to be purchased by such users, distribution of the mission planning software would be greatly hindered.

Since 1986, the development of Silvabase and the conversion of mission planning software from MIPS has been ongoing. Boeing Computer Support Services (BCSS), a programming support contractor at MSFC, has performed the task of designing, coding, and documenting Silvabase according to requirements established by the Mission Operations Laboratory. Silvabase owes its existence to the bright team of programmers at BCSS.

A USER'S PERSPECTIVE

What is it?

Silvabase consists of a flexible file format, a library of subroutines for reading and writing Silvabase files, a utility program for interactively manipulating Silvabase file contents, and substantial documentation. Silvabase is implemented on Digital Equipment Corporation's VAX 11 family of computers using the VMS operating system in VAX-11 FORTRAN. Currently, Silvabase is only being utilized from FORTRAN programs, but it should be useable from programs in other languages that adhere to the VAX subroutine calling standards. The name "Silvabase" is created from the Latin word "Silva" meaning forest. As is described below, Silvabase files are based largely on the B+ tree data structure, and thus the concept of a forest is brought to mind.

Silvabase is a data file management system that is particularly suitable for storage of small to medium-large amounts (up to 2.1 gigabytes) of data in which the principal data is organized into collections of homogeneous records using a single key. The key is the first field in a data record with a unique value and is used to locate that record. Silvabase provides for the key to be one of nine different data types, from integer to a string of eight characters. Other data items in Silvabase may be one of twenty two different data types. Multiple collections of records, called "subjects" may exist on one Silvabase file at the same time. Subjects may contain data items outside of and associated with the data records, which are normally viewed as constant with respect to the records. Files may contain data items outside of all the subjects which are viewed as constant across all subjects. Exactly what file variables exist on a given file, what subjects are included in a given file, what subject variables appear on each subject, and the structure of the data records on each subject is left entirely up to the applications programs which write the file and subjects. Although most of the capabilities of Silvabase were created with mission planning requirements in mind, this flexibility and tiered structure of the data makes Silvabase potentially applicable to a very wide range of data storage needs.

Because of the requirement that mission planning software be easily distributed, Silvabase was created with portability between VAX/VMS systems in mind. There is nothing in Silvabase that hinders its implementation on another VAX running the VMS operating system. Silvabase uses standard VAX file names. The files may be accessed on a remote node via network connections. And Silvabase does not add to or restrict the standard VAX file protection, expiration or automatic deletion functions.

A User's Concept of a Silvabase File

Figure 1 shows a conceptual diagram of a Silvabase file illustrating the components as seen from the viewpoint of a file user. These are described below.

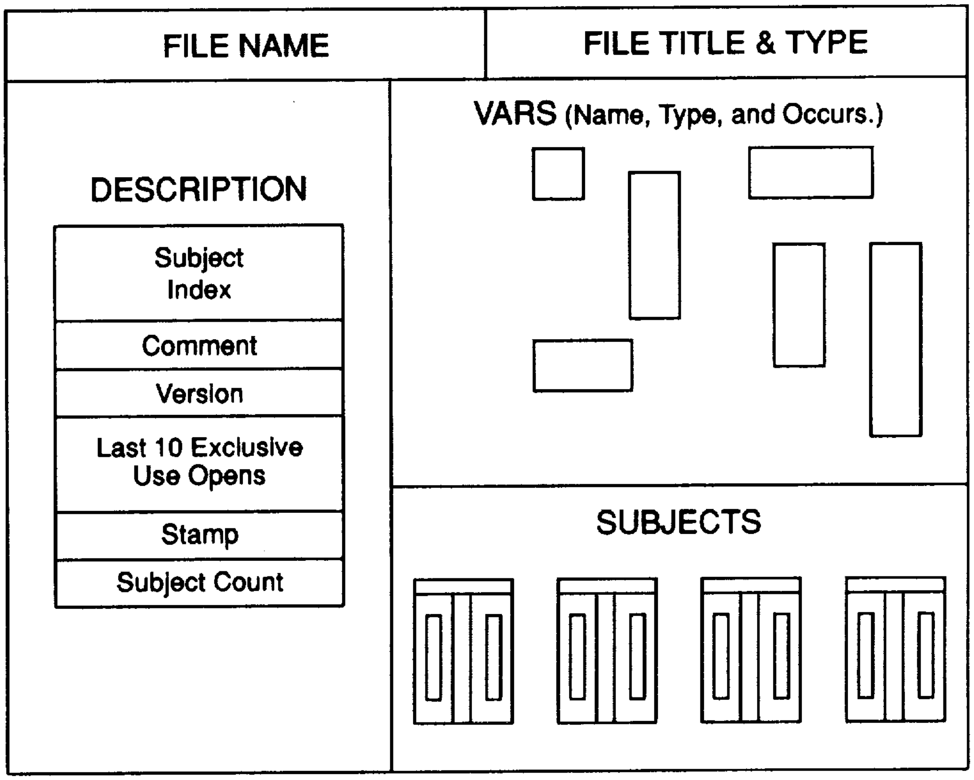


Figure 1: **Components of a Silvabase File**

- FILE NAME - This is the VAX/VMS name of the file. It is not actually contained in the Silvabase file.
- FILE TITLE - This is an optional 72 character title that may be written to the file.
- FILE TYPE - This is a four-character identifier which would be set to a predefined value and is intended to indicate that the file has certain standard contents to any user who may read the file.
- SUBJECT INDEX - This is an ordered list of the names of all of the subjects that are contained in the file. This list may be reordered by the user to suit his requirements.
- COMMENT - The file comment is a 216-character long string that may contain whatever description the user chooses to write.
- VERSION - This field is set by the Silvabase software and indicates the version of Silvabase software used to create the file.
- LAST 10 EXCLUSIVE USE OPENS - Silvabase keeps track of the times that a file was opened for "exclusive use," i.e. with intent to write. Silvabase maintains this list which contains the dates and times of the last 10 exclusive use opens.
- STAMP - The stamp is the so called "certification stamp." This field may be used by a user in authority to indicate that the file and data it contains is official or verified. This field contains the name of the certifying authority and the date and time of the certification. If the file is opened with the intent to write, it is automatically decertified.

- SUBJECT COUNT - This is simply the number of subjects on the file and equals, naturally, the number of subjects in the subject index.

- VARIABLES - The file variables, as mentioned above, are simply any number and type of data items which the user creates and assigns values to. These are intended to contain data which is descriptive of the entire file. The file variables may be of any of the twenty-two data types available for variables in Silvabase. Also, these variables may have multiple occurrences, i.e. they may be dimensioned as one dimensional arrays.

- SUBJECTS - Silvabase subjects are the principal structure for containing data in a Silvabase file. Each subject is independent of the others. The structure of a subject is definable by the user. The features of a Silvabase subject are described below.

A User's Concept of a Silvabase Subject

Figure 2 shows a conceptual diagram of a Silvabase subject illustrating the components as seen from the viewpoint of a Silvabase file user. Each component of a subject is described below.

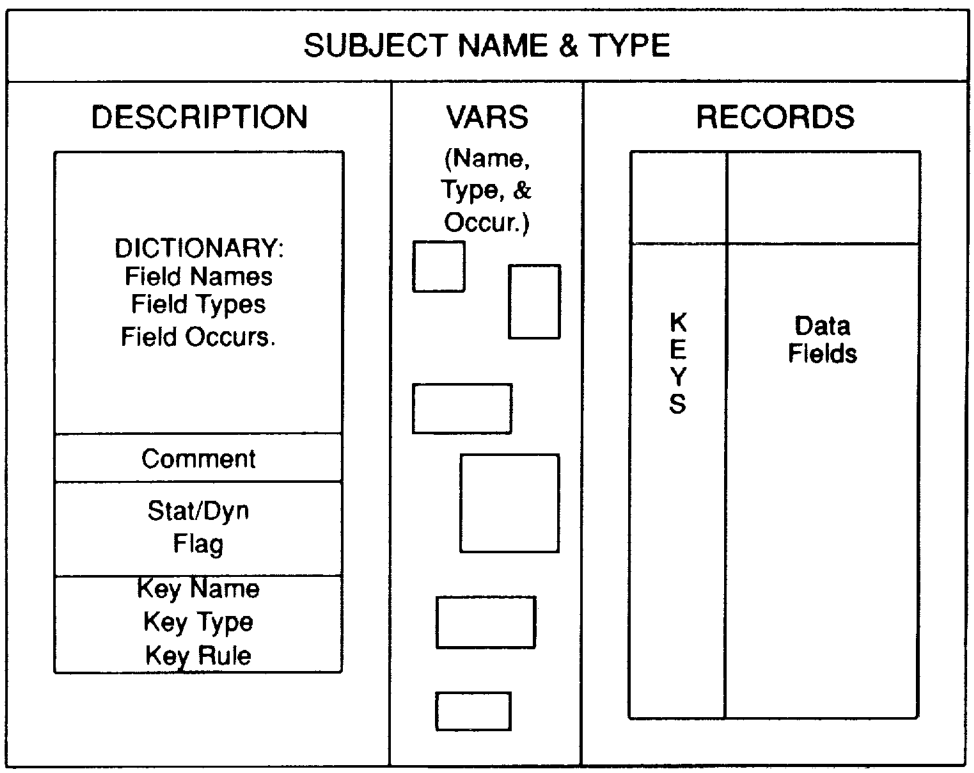


Figure 2: **Components of a Silvabase Subject**

- SUBJECT NAME - Every Silvabase subject has a unique name up to sixteen characters.

- SUBJECT TYPE - This is a four-character identifier which would be set to a predefined value and is intended to indicate that this subject has certain standard contents to any user who may read the subject. It is to the subject what the File Type is to the file.

- DICTIONARY - This is a listing of the names, data types, and number of occurrences (dimension) of all the data fields stored on the subject records. The definition for the key is stored separately (see below) and may be retrieved separately.

- COMMENT - The subject comment is a 72-character long string that may contain whatever description the user chooses to write to describe a subject.

- STATIC/DYNAMIC FLAG - This flag indicates whether this subject is considered "static" or "dynamic." A "static" flag on a subject will prevent that subject from having records inserted or deleted. They may be appended or edited. The flag is intended as a safety feature and may be changed at will.
- KEY NAME, TYPE AND RULE - This information describes the key field of the data records. The key rule is currently only applicable to interval keys and describes how those intervals may relate to each other. Key Rule 1 does not allow the intervals to share a common endpoint. Key Rule 2 allows this. All interval keyed subjects in mission planning are currently defined to use Key Rule 2.
- VARIABLES - The subject variables are simply any number and type of data items which the user creates and assigns values to. These are intended to contain data which is descriptive of this entire subject. These variables may be of any of the twenty two data types available for variables in Silvabase. Also, these variables may have multiple occurrences, i.e. they may be dimensioned as one dimensional arrays.
- RECORDS - The records in a subject are designed to be the principal data structure in a Silvabase file. The records may be conceptually described as row after row of data items, each row having the same structure as the others. Individual records in a Silvabase subject may be read, written or deleted sequentially or randomly. Of the data items in a record, the first one is considered the key field and the others are referred to as data fields. On each record, the key must have a unique value that differentiates that record from the others. The records are stored in ascending order of the key and this order is maintained by Silvabase. The data fields on each record then contain values that logically correspond to the key. For example, in mission planning, one type of Silvabase subject contains a representation of time in the key and the data fields contain information describing the position and velocity of the Space Shuttle in orbit. This type of subject then comprises a time history of the Shuttle's location where each record represents another moment in time.

The key of a Silvabase subject may be one of nine allowed data types, and each data field may be one of the twenty two data types Silvabase allows. How a subject is built, i.e. what type of key is used, what types and how many data fields are used on each record, and what each data item represents, is what makes a given Silvabase subject a unique subject "type." The example described above is a trajectory type of Silvabase subject.

As can be seen from this description of a subject, there are myriad ways of defining a Silvabase subject by selecting a type of key and a number and type of subject variables and data fields. In mission planning, dozens of different standard subject types have been defined, each for storing a different kind of data set such as a trajectory, vehicle attitudes, and a payload's use of resources.

The Tools of a Silvabase User

Naturally, to create and manipulate a Silvabase file, the user will need some tools and accessories. The most crucial tool is of course the Silvabase Library, a library of VAX FORTRAN subroutines which constitute all the primary functions one would need to perform on a Silvabase file, from creation to field by field editing. This library, along with detailed user information, is documented in The MASE/Silvabase Programmer's Guide [2]. (MASE is the Marshall Applications Support Environment, a BCSS developed set of programming tools of which Silvabase is one.)

Another important tool for Silvabase file users will be "SUP", the Silvabase Utilities Program. This is a program still under development which will give file users access to file manipulation functions in an interactive setting. SUP will allow users to interactively create files and subjects, enter and edit data, plot and tabulate data, and perform scripted file operations. Currently, development of SUP is in the requirements definition stage, but a predecessor to SUP called ISUP (Interim SUP) has already been developed. ISUP has many, but not all, of the functions planned for SUP. Due to the great flexibility of Silvabase and the special needs of mission planning, both ISUP and SUP will address mission planning requirements and will not be capable of performing every Silvabase function on every conceivable type of Silvabase subject. Like ISUP, SUP will also include a documented library of callable functions which are features of the program above and built from the primary Silvabase functions [3]. SUP will also be documented with a user's guide as is ISUP [4].

As mentioned before, mission planning has defined a couple of dozen standard Silvabase subject types specific to mission planning. These definitions have given rise to three other accessories that make Silvabase programming easier in the mission planning world. First of all, the Silvabase subject types for mission planning are described in a document that details the subject user type, subject variables, key type, data fields and other aspects of each subject type [5]. Along with that, a Silvabase file has been established which contains one empty example of each standard subject type. These subjects can be used by Silvabase as templates any time a user needs to create another subject of a standard type. Finally, a text library containing fragments of FORTRAN code was created. This library is referred to as the structures library. Each entry in the library contains the FORTRAN declarations necessary to create a data structure suitable for containing the record from one of the standard subject types. Programmers working on applications that must read or write standard mission planning Silvabase subjects can use these code fragments in their programs to quickly and easily create correct data structures for holding records of standard subjects. This simplifies the coding job and insulates application programs from potential changes in standard subject definitions.

One final Silvabase accessory that is not as important to users of Silvabase files is the MASE/Silvabase Internals Document. This document was written by BCSS developers of Silvabase for themselves. It details the internal workings of Silvabase files and how the library software operates on them.

Special Functions

The Silvabase software includes a number of special functions that are required to do the types of data manipulation needed in mission planning. The following are some of those functions that have been built into Silvabase.

- Key Offset — This function will offset each key in a specified subject by a specified amount. It will not affect the position of any key within the subject.
- Data Field Offset — This function will offset a numeric data field by a specified amount.
- Complement — This function stores time intervals not included in a specified subject on an output subject. The intervals to be complemented are within a specified domain of values.
- Union — This function will union the intervals of two or more interval keyed subjects. The results are stored on a new subject whose intervals contain the times included in any of the input subjects' intervals (logical 'OR').
- Intersect — This function will intersect intervals of two or more subjects. The results are stored on a new subject whose intervals contain the times included in all of the inputs subjects' intervals (logical 'AND').
- Intersect and Transfer Data Fields — This function will perform the same operation as the previous intersect function with the additional feature that specified data fields can be transferred to the new subject.
- Interpolate — This function uses the central difference formula to interpolate two numeric data fields on a key value.

INTERNAL IMPLEMENTATION

B+tree

Silvabase is designed to allow sequential as well as random access to records. Sequential access is very desirable because it eliminates the need for disk seeks when reading or writing sequential records. Random access is needed for searches, insertions, and deletions. In order to accomplish these operations efficiently, Silvabase uses a B+ tree for each subject on the file. The B+ tree is a B-tree combined with a sequential set of data records called a sequence set.

The sequence set is extremely efficient for sequential access and appending new records. The set is made up of physically contiguous records organized in sequential order by key. The records are organized into blocks. The blocks are linked together sequentially according to the range of key values contained within each. Each block points both to the next and previous block in the set. This blocking of records allows for easy maintainence of the sequence set because it limits the effects of insertions and deletions to the records within or near the block containing the change. Using sequence blocks allows several sequential records to be read into memory at once rather than reading them one at a time. The file is read sequentially by loading a block of records into memory and reading each record until the last record in the block is reached. The next block is then located using a pointer and loaded into memory for further reading. This greatly lessens the amount of disk seeks needed for transferring records to or from disk. Sequential search performance is also improved since the records can be searched in memory instead of on disk.

Records may be appended, inserted, or deleted in the sequence set. Appending to the end of the sequence set is very straightforward and can be done with tremendous effectiveness. Insertions into a sequence block are made by first performing a binary search for the correct position within the block for the new record. The records are then shifted to make room for the new record and it is inserted. If overflow occurs, the block is split into two blocks and the new block is linked into the list. Deletions are made using simple collapse procedures and concatenating when a block underflows. These operations keep the sequence set in order and eliminate the need for sorting.

A B-tree is used as an index to organize the sequence blocks. It provides fast, efficient random access to records and is completely maintainable. The nodes of the B-tree contain separators. These separators are not the actual keys of the records but indicate the range of keys located within a particular sequence block. Several separators along with child pointers are stored in one node in the form of an ordered dense list. This allows several separators to be read into memory at once. This also simplifies making changes to the node, allows binary searching, and eliminates disk accesses. The same advantages apply here as apply to blocking records for the sequence set. The B-tree is constructed from the leaves up, which keeps the tree well-balanced for efficient, fast searching. Locating a specific record involves descending through the B-tree, loading a node into memory, and performing a binary search on the separators within the node to find the path to the sequence block containing the record. A stack is used to keep track of the path through the B-tree. Once the correct sequence block is located, a binary search is performed on the keys within the block. If the record is very small, the data is stored within that block. Otherwise, the record will have a pointer to the location of the desired data. Even though all searches must descend through the entire tree to the sequence blocks, the performance is so good for the worst case search that this is not a concern.

Chunks

The size of sequence blocks and B-tree nodes is the same. They each fill one physical record of 1024 bytes. Silvabase uses block I/O, since records on the VAX are written to and read from the disk 512 bytes at a time, the physical records must be a multiple of 512 bytes in size. Small records increase I/O and large records require more space. This size of 1024 bytes is large enough to decrease I/O but small enough for good space utilization.

A physical record for a Silvabase file is called a "chunk". When a Silvabase file is created, several chunks are allocated, some for specific potential purposes. Each chunk has a chunk header. This is used used to save information about the chunk and to link it both forward and backward with the other chunks in the file. A chunk is referenced by its sequence number in the file. The first chunk is always the file header.

Structures

The internal organization of a Silvabase file involves the use of several types of internal structures. Each structure is stored in a separate chunk. Some structures require several chunks and are linked together in a "chunk chain". The internal structures of a Silvabase file include the following:

- FILE HEADER — This keeps information about the file such as its title and type. It also keeps track of the chunk numbers of the other internal structures in the file.
- FILE VARIABLES — As defined previously, these are used to store any type of information that pertains to all the subjects on the file.
- SUBJECT VARIABLES — As defined previously, these are used within each subject to store any type of information that is constant throughout the subject.
- SUBJECT INDEX — As described earlier, this contains a list of all the subjects on the file in order. It stores the name, type, and subject description pointer for each subject. A subject may be found using a sequential search for the subject name or by its sequence number in the list.
- SUBJECT DESCRIPTORS — There is one subject description for every subject on the file. A subject description serves as the header record for a subject in the same way that the file header does for the file. The descriptions are kept separate from the subject index for faster searching of the index.
- SUBJECT COMMENTS — Each subject may have a comment associated with it. As defined above, this comment is pointed to by the subject description.

- DICTIONARIES — This stores the format of the record structures in the file. It contains each field name, type, and occurrence.
- B+TREE INDEX SET — This stores the nodes of the B-tree.
- B+TREE SEQUENCE SET — This stores the sequence blocks for each subject. There are 5 types of sequence blocks depending on the size of the key and data to be stored. In many cases, a pointer to the data is saved with the key instead of the data itself.
- DATA BLOCKS — This stores the actual data for the record fields for each subject.
- EMPTY CHUNKS CHAIN — This is a set of empty chunks linked together and available for use by Silvabase.

When a Silvabase file is opened, parts of the file header are read into a File Information Table (FIT) which is located in memory and kept for fast access to the structures within the file. The Subject Information Table (SIT) serves the same purpose for an open subject on the file.

CAPABILITIES

Performance

Silvabase traverses through the B-tree using a binary search on each node to locate keys. This technique eliminates half of the remaining keys in the subject with every comparison of keys. The length of a worst case binary search is calculated using the following formula:

$$W(M,N) = 1 + \frac{ln(N+1)/2}{ln(M/2)}$$

M is the B-tree order - maximum number of children possible for each node
N is the number of keys in the tree
This formula assumes that each node is only half full.

Both the order of the B-tree and the number of records within the subject effect the amount of disk accesses needed for the tree traversal. The rate of additional disk accesses becomes smaller as more records are added. A large B-tree order also reduces disk accesses but must be limited for maximum gains. These effects are shown in Figure 3 and Figure 4.

Capacities

Most of the limitations on Silvabase were chosen in order to accomadate the needs of Mission Planning. Some restrictions of a Silvabase file are the following:

- File Size — This limit exists because Silvabase allows 21 bits for physical record representation. Since the physical record length is 1024 bytes, the maximum number of chunks must be 2^{21} =2,097,151 or 2.1 gigabytes per file.
- File and Subject Variables — There is no set limit on the number of file or subject variables allowed.
- Subjects — The maximum number of subjects that may be on the file is 5000.
- Records — There is no restriction on the number of records per subject, but the number of data fields per record cannot exceed 1000. The maximum size of a record is 8000 bytes.
- FIT and SIT — As many as 12 files and 48 subjects may be open at one time.

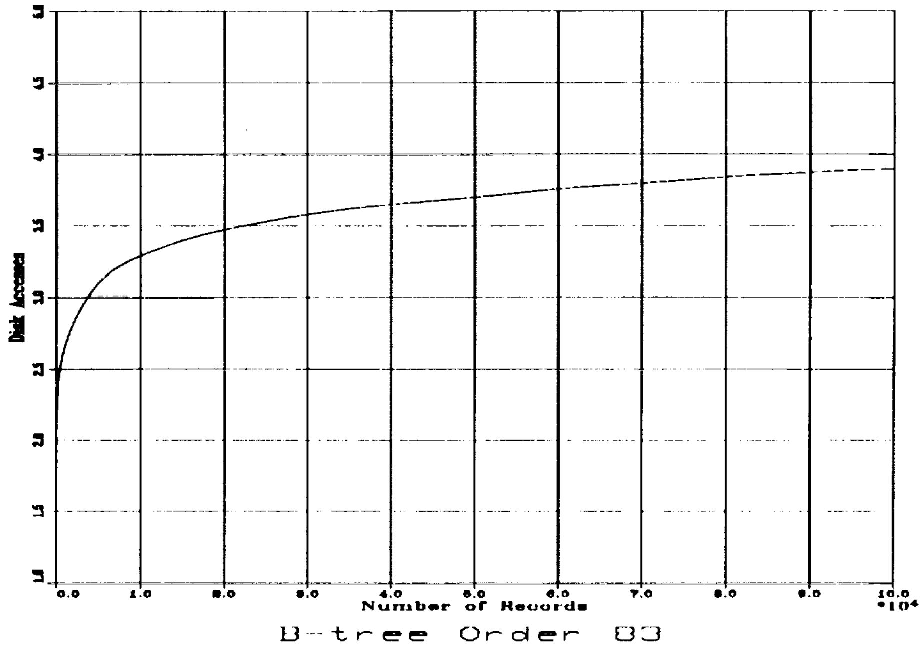


Figure 3: At Constant B-Tree Order 83

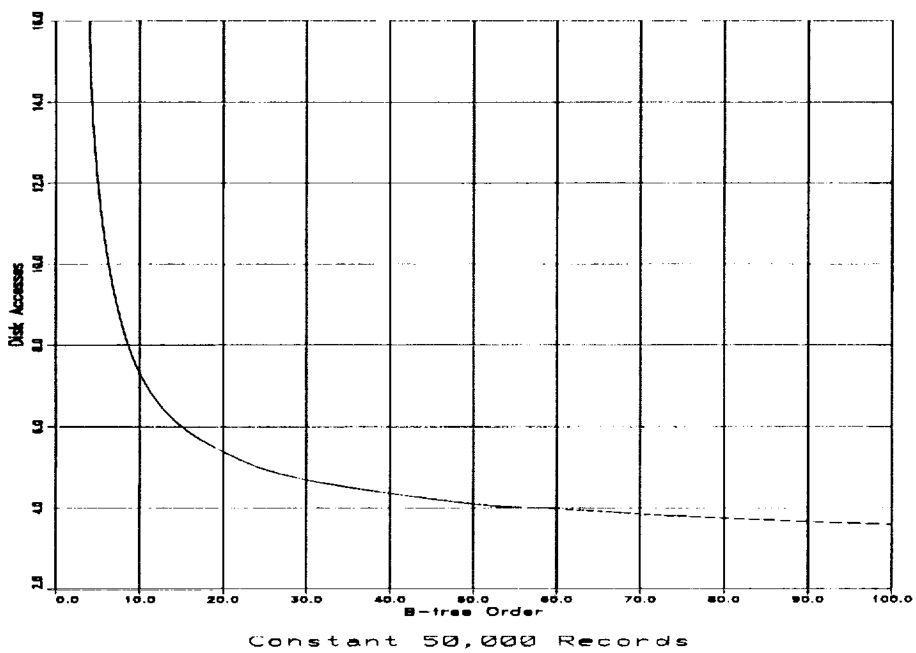


Figure 4: At Constant 50,000 Records

SUMMARY

Silvabase was created out of a need in the Mission Operations Laboratory for a powerful, maintainable data file management system that both addressed specific mission planning needs and was flexible enough to adapt to the ever changing requirements in mission planning. Silvabase offers the data-generating user a tiered structure of records within subjects within files, numerous data types, and the flexibility to organize these in whatever way best suits his or her needs. It is based on the powerful B+ tree data structure that gives file accessing its efficiency. This combination of features makes Silvabase an attractive data file management system for innumerable applications.

REFERENCES

[1] S. M. Spillers. *The Future of VAX MIPS*. Boeing Computer Support Services, October 17,1986.

[2] *MASE/Silvabase Programmer's Guide*.Boeing Computer Support Services, March 29,1991.

[3] *ISUP Interim Silvabase Utilities Program Library User's Guide*. Boeing Computer Support Services, July 23,1990.

[4] *ISUP Interim Silvabase Utilities Program User's Guide*. Boeing Computer Support Services, July 2,1990.

[5] S. J. Lambing. *Formats for Mission Planning Subject Types in Mase Silvabase Files*. National Aeronautics and Space Administration, MSFC, August 9, 1990.

ELECTRO-OPTICS

(Session C2/Room A1)

Wednesday December 4, 1991

- **Nonlinear Optical Polymers for Electro-Optic Signal Processing**
- **High-Resolution Optical Data Storage on Polymers**
- **Laser Discrimination by Stimulated Emission of a Phosphor**
- **Pulsed Laser Prelasing Detection Circuit**

N92-22476

NONLINEAR OPTICAL POLYMERS FOR ELECTRO-OPTIC SIGNAL PROCESSING

Geoffrey A. Lindsay
Polymer Science Branch Head, Code 3858
Chemistry Division, Research Department
Naval Weapons Center, China Lake, CA 93555

ABSTRACT

Photonics is an emerging technology, slated for rapid growth in communication systems, sensors, imagers and computers. Its growth is driven by the need for speed, reliability and low cost. New nonlinear polymeric materials will be a key technology in the new wave of photonic devices. Electron-conjugated polymeric materials offer large electro-optic figures of merit, ease of processing into films and fibers, ruggedness, low cost and a plethora of design options. Several new broad classes of second-order nonlinear optical polymers have been developed at the Navy's Michelson Laboratory at China Lake, California. Polar alignment in thin film waveguides was achieved by electric-field poling and Langmuir-Blodgett processing. Our polymers have high softening temperatures and good aging properties. While most of the films can be photobleached with ultraviolet (UV) light, some have excellent stability in the 500 - 1600 nm range, and UV stability in the 290 - 310 nm range. The optical nonlinear response of these polymers is subpicosecond. Electro-optic switches, frequency doublers, light modulators and optical data storage media are some of the device applications anticipated for these polymers.

INTRODUCTION

Recently, there has been immense interest in second-order nonlinear optical polymer (NLOP) films[1] because of their large nonlinear optic coefficients, and the ease of casting thin films on many substrates. NLOP can be spin-cast into optical waveguides using conventional microlithographic equipment. These new polymers are under development for applications in electro-optical modulators,[2] switches,[3] waveguides[4] and optical interconnects.[5]

Electric-Field Poling. A useful process for imparting the second-order NLO property is electric field poling.[6] A polymer containing dyes (i.e., chromophores) which have large dipole moments,[1a] is heated and poled near the glass transition temperature (Tg), then cooled below the Tg while still applying the electric field. After the external field is removed, a net alignment of dipole moments can remain essentially locked in the film for years as long as the temperature of the film remains well below any solid state transition, such as the glass-rubber transition (Tg). This imparts a noncentrosymmetry to the film which is necessary for frequency doubling (i.e., the thin film's ability to generate second harmonic light when a laser beam is transmitted through it)[7], and the Pockels effect (i.e., its ability to change index of refraction as a function of an electric field applied across the film).

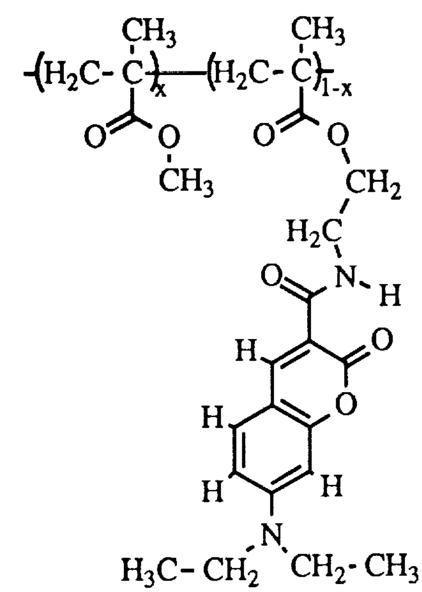
Langmuir-Blodgett Processing. Another technique for fabricating polar films is Langmuir-Blodgett (LB) processing. There are several good reference books on LB technology.[8] Simply stated, an organic compound is floated on a liquid, usually water, and a solid substrate is dipped through the air-water interface depositing a single molecular layer per stroke on the substrate. Eastman Kodak researchers have built up a micron-thick, NLOP film of optical high quality.[9] Turn-key, computer-automated LB troughs are available from many commercial suppliers. Appendix I briefly reviews LB film deposition. The Naval Weapons Center chemistry laboratory has a NIMA LB trough equipped with two compartments such that multilayer films, alternating $(AB)_n$ times, can be deposited automatically.

Developments in the NLOP Field. The development of nonlinear optical polymers has been rapidly evolving over the last eight years. The earliest materials investigated were the guest-host systems comprised of a dye (chromophore) dissolved in a glassy polymer matrix, such as Disperse Red #1 dissolved in poly(methyl methacrylate).[10] These mixtures exhibit second-order optical nonlinearity, but it slowly decays over a period of months. Soon investigators found that if the dye is chemically attached to the polymer backbone, the stability of the poled films is greatly enhanced. Many laboratories have reported novel sidechain NLOP compositions[11] in which the chromophores are attached at one site, pendent to the polymer backbone. The sidechain polymers are easy to process, and most of the research effort has focused on this class of NLOP. Attaching chromophores along a polymer backbone so that they form the backbone of the polymer is another interesting configuration which may yield processable polymers with a higher concentration of chromophoric material (the active nonlinear optical component). Developments in each of these types of NLOP will be described in the following sections. Another approach which will not be covered are the crosslinked NLOP. This class of NLOP are formed by carrying out a chemical reaction (the crosslinking reaction) in the presence of an electric field. Crosslinked NLOP promise to yield films which are more thermally stable; however, in practice, they are very difficult to process into films which have low scattering losses

THREE CLASSES OF NONLINEAR OPTICAL POLYMERS

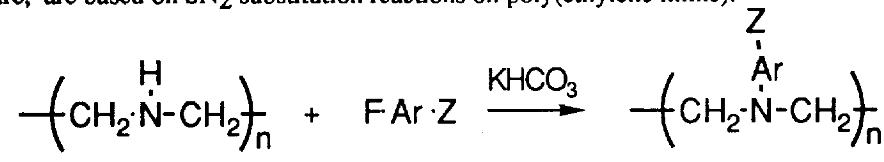
1. Sidechain NLOP. The first sidechain polymers made in our laboratory were designed for Langmuir-Blodgett (LB) deposition.[12] These polymers were based on the stilbazolium chromophore which has one of the largest second-order nonlinear coefficients due in part to the formal positive charge on the picolinium group. Due to migration of the charges, these polymers can not be aligned by electric field poling, but LB deposition gives a high degree of polar order. Thin LB films of these polymers can be easily patterned when exposed to UV light to form optical gratings and holograms. A patent application has been filed.[13]

Quite a large number of sidchain polymers were developed in our laboratory which have no charges, hence are ideal candidates for electric-field poling.[14] Most of the sidechain polymers developed in our laboratory are described in references 11e through 16. One of the sidechain polymers, with which we have the most experience, is based on the coumarin chromophore:



This NLOP is quite fluorescent and very photo stable. The damage threshold at 532 nm is about 60 GW/cm^2 for short pulses.[14c] We have prepared a sidechain polymer with a glass transition temperatures of 170° C, and polymers with even higher thermal stability are under development. A patent application has been submitted.[15]

Several new sidechain polymers under development in our laboratory, which have not yet been reported in the scientific literature, are based on SN_2 substitution reactions on poly(ethylene imine):



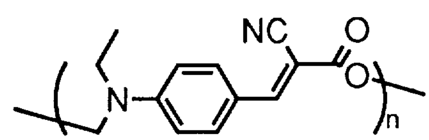
where FArZ is a group containing an aromatic ring system with at least one electron accepting group (Z) attached para to fluorine. Hence, a large dipole moment results because the amine is an excellent electron donating group. Since the amine lies in the backbone, a high density of chromphores is possible.

Another potentially useful chromophore, 4-nitrocinnamylidene, developed by Prof. Sam Huang at the Univ. of Connecticut, was converted to a novel sidechain NLOP by our group. The nitrocinnamylidene chromophore is nearly "water white" in the visible region of the spectrum.

We have been designing synthetic polypeptides which can form ordered monolayers at the air-water interface. We have synthesized novel, chromophore-substituted polypeptides designed for self-organization (in beta-sheet conformations). These polymers are prepared by derivatization of preformed polypeptides, and by polymerization of derivatized amino acids. An invention disclosure has been submitted.[16]

Sidechain NLOP are very attractive for many applications, and still hold the record for degree of alignment and NLO figures of merit. They also may be crosslinked by various techniques to give higher stability.

2. Isoregic Mainchain NLOP. The configuration of chromophores attached head-to-tail along a polymer backbone is called the isoregic mainchain configuration. Isoregic NLOP have been of interest, in part, because it was proposed that for chemically connected dipoles compared to the unconnected case, the propensity to align in an electric field is greatly enhanced due to larger molecular dipole moments.[17] One example of an isoregic polymer prepared in our laboratory is shown below[18]:



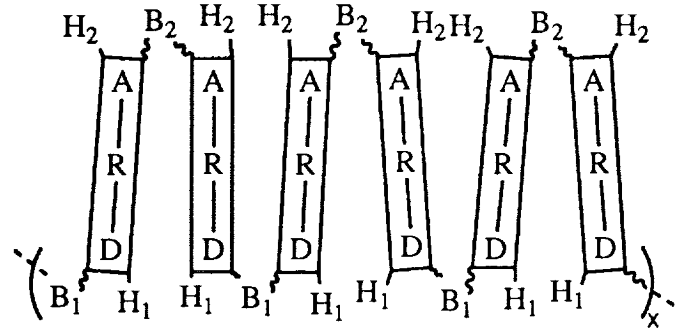
This polymer is completely soluble in common solvents, is amorphous, has a glass transition temperature of 105° C, and can be polarized by electric-field poling.

The isoregic configuration offers the possibility of aligning the chromophores by mechanically stretching the polymer. However, one must find a technique to induce all of the dipoles in the mainchains to point in the same direction, or else they will cancel each other even if the chains are aligned.

We remain bullish about the possibilities for isoregic NLOP. We describe our progress in publications listed in reference (18). We have submitted a patent application.[19] Our current interests in this area, in collaboration with other laboratories, include: a) poling in a supercritical fluid to plasticize the polymers (D. Soane at U.C. Berkeley), and 2) growing polymers from a surface in one direction to effect a perfect polar alignment (C. Martin at U. Colorado).

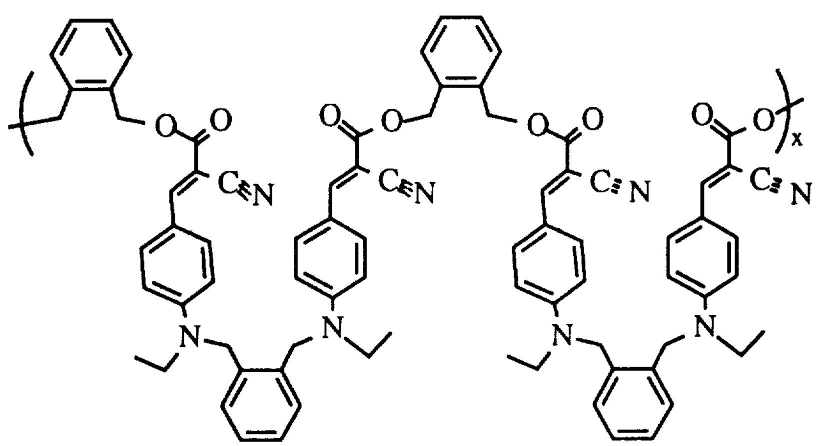
3. <u>Syndioregic Mainchain NLOP</u>. Our laboratory was the first to explore the head-to-head (syndioregic) configuration. If the syndioregic backbone is viewed in the extended conformation, the dipole moments appear to cancel each other out. However, we felt it might be possible to cause the backbone to wrinkle into a rigid, folded conformation by selecting proper groups for local molecular interactions.

It is clear from molecular models that one can place bridging groups (B) and other "helper" groups (H) between rigid, syndioregic chromophores (ARD) that will encourage the backbone to fold.



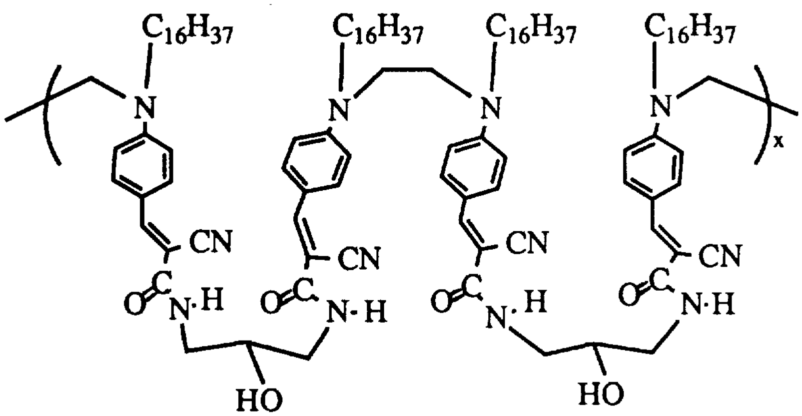
There would be two molecular processes involved in forming the final solid state conformation. One process would be the folding of the chromophoric groups relative to their immediate neighbors along the backbone, and the other process would be the placing of the entire polymer chain relative to neighboring polymer chains. The folded polymer may act as a rigid ensemble whose dipole moment is the sum of the constituent chromophores. However, during electric field poling, the folding of the chromophores should also be influenced by the applied electric field.

We are now in the process of synthesizing new syndioregic polymers with rigid bridging groups and high glass transition temperatures, such as the polymer shown below:



Preliminary samples of this polymer have glass transition temperatuers of about 150° C, and are soluble in many common solvents. Electric field poling of thin films is now underway. Eliminating the methylene groups on the xylyl bridges will further increase the glass transition temperature.

By placing hydrophilic and hydrophobic groups in the right places, one can also design syndioregic polymers for LB deposition:



We have made these and other syndioregic polymers films which have nonlinear optical properties (as measured by second harmonic generation).[20] Several of these polymers can be patterned by UV etching/bleaching. The patent application is in process.[21]

KEY PROPERTIES AND APPLICATIONS

Temporal Response. The response time of NLOP has been shown to be sub-picosecond.[14c] This is essential for rapid switching in hybrid electro-optic computers, and sensor protection against lasers. We have identified many applications for missile systems involving signal modulation and laser beam deflection.

Holographics. NLOP can be processed on conventional photolithography equipment. While most of the NLOP films can be photobleached with ultraviolet (UV) light, some have excellent stability in the 500 - 1600 nm range, and UV stability in the 290 - 310 nm range. Chromophores can be tailor-made to absorb light in certain regions of the spectrum, and to be transparent in others. This means that a grating can be etched by exposing the NLOP through a mask in the absorbing region, while the grating spacing is designed to diffract light in a nonabsorbing region. The concept of three dimensional images stored in high density in NLOP has not yet been developed, but it offers many exciting possibilities, such as electro-active (switchable) holograms.

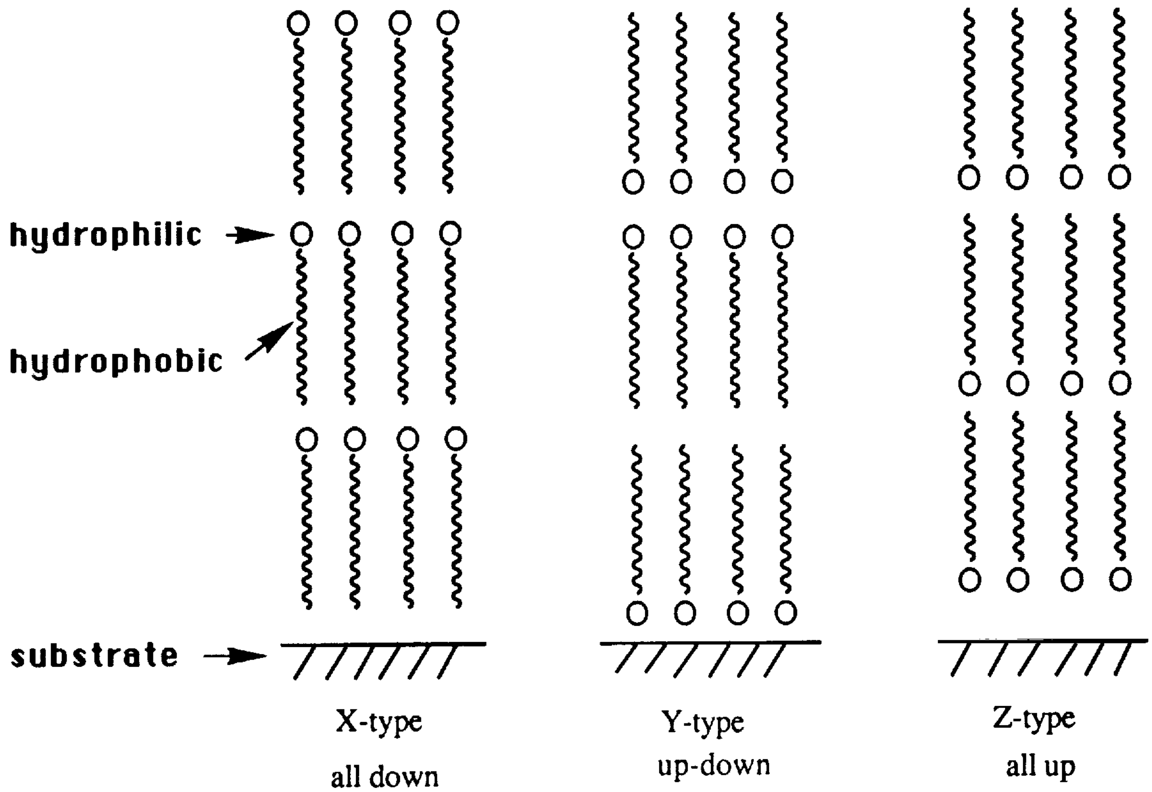
Mach-Zehnder Interferometer. A major applications focus for NLOP materials has been optical signal modulation. There have been several reports of modulation rates in the range of 20 to 40 GHz. NLOP have the advantage of low dielectric constants and high electro-optic coefficients compared to conventional materials such as lithium niobate.

ACKNOWLEDGEMENT

The author gratefully acknowledges the sponsorship of the Office of Naval Research, and his research colleagues at the Naval Weapons Center who prepared and characterized the nonlinear optical polymers: Dr. John Fischer, Dr. Ronald Henry, Dr. James Hoover, Dr. Robert Kubin, Dr. Michael Seltzer and Dr. John Stenger-Smith.

APPENDIX I. Langmuir-Blodgett (LB) Film Deposition

Multilayer LB films can be formed in three different configurations, as depicted on the next page. Historically, these are called "X"-, "Y"-, and "Z"-type films, where X is made by depositing always on the down-stroke, Z is made by depositing always on the up-stroke, and Y is made by alternating up- and down-strokes. For the case in which the large dipole moment of the sidechain chromophore is normal to the polymer backbone (and the backbone is in the plane of the air-water interface), all-up or all-down films will be polarized, and the up-down films will not be polarized (dipoles in adjacent layers cancel out). The Y configuration is thermodynamically more stable; and, sometimes X and Z configurations spontaneously rearrange in the solid state to form the Y configuration. Hence, one may interleave a sidechain chromophoric polymer with an optically inert spacer layer (having little dipole moment) and arrive at a stable, polarized film in the Y configuration.



---

1 a) D. J. Williams, *Angew. Chem. Int. Ed. Engl.*, **23**, 690 (1984);
b) D.S. Chemla and J. Zyss, Eds., Vol. 1, *Nonlinear Optical Properties of Organic Molecules and Polymeric Materials* (Academic, New York, 1987);
c) A.F. Garito, K.Y. Wong and O. Zamani-Khamiri, in *Nonlinear Optical and Electroactive Polymers*, P.N Prasad and D. R. Ulrich, Eds. (Plenum, 1988).

2 G.H. Cross, A. Donaldson, R.W. Gymer, S. Mann, N.J. Parsons, D.R. Haas, H.T. Man and H.N. Yoon, *SPIE Proc.* **1177**, 79 (1989).

3 P. Kaczmarski, J.-P. Van de Capelle, P.E. Lagasse, R. Meynart, *IEE Proc.* **136, Pt. J** (3), 152 (1989).

4 M.J. McFarland, K.K. Wong, C. Wu, A. Nahata, K.A. Horn, J.T. Yardley, *SPIE Proc.* **993**, 26 (1988).

5 C.A. Eldering, S.T. Kowel, P. Brinkley, N. Matloff, T. Schubert, R. Gosula, *SPIE Proc.* **1151**, 72 (1989).

6 a) K.D. Singer, M.G. Kuzyk, W.R. Holland, J.E. Sohn, S.L. Lalama, R.B. Comizzoli, H.E. Katz, M.L.Schilling, *Appl. Phys. Lett.* **53**(19), 1800 (1988);
b) M.A. Mortazavi, A. Knoesen, S.T. Kowel, B.G. Higgins, A. Dienes, *J. Opt. Soc. Am. B* **6**(4), 733 (1989);
c) C.P.J.M. van der Vorst, S.J. Picken, *J. Opt. Soc. Am. B*, **7**(3), 320 (1990).

7 Y.R. Shen, "The Principals of Nonlinear Optics," John Wiley & Sons, NY, 1984.

8 a) G. L. Gaines, *Insoluble Monolayers at Liquid-Gas Interfaces*, Interscience: New York, 1966;
b) G. G. Roberts, *Langmuir-Blodgett Films*, Plenum Press: New York, 1990.

9 T. Penner, et al., *SPIE Proceedings*, **1064**, San Diego, CA, July 25, (1991).

[10] K.D. Singer, J.E. Sohn, and S.J. Lalama, *Appl. Phys. Lett.* **49**, 248 (1986).

[11] a) E.W. Choe, U.S. Patent 4,755,574, Jul. 5, 1988;
b) R. N. DeMartino, U.S. Patent 4,757,130, Jul. 12, 1988;
c) P. LeBarny, et al., *SPIE Proc.* **682**, 56 (1986);
d) A.C. Griffin, *SPIE Proc.* **682**, 65 (1986);
e) T.J. Marks, U.S. Patent 4,935,292, Jun. 19, 1990;
e) K.D. Singer, M.G. Kuzyk, W.R. Holland, J.E. Sohn, S.L. Lalama, R.B. Comizzoli, H.E. Katz and M.L.Schilling, *Appl. Phys. Lett.*, **53**(19), 1800 (1988);
f) L..M. Hayden, G.F. Sauter, F.R. Ore, P.L. Pasillas, J.M. Hoover, G.A. Lindsay and R.A. Henry, *J. Appl. Phys.*, **68**(2), 456 (1990).

[12] a) "New Hemicyanine Dye-Substituted Polyethers as Nonlinear Optical Materials," R. C. Hall, G. A. Lindsay, S. T. Kowel, L. M. Hayden, B. L. Anderson, B. G. Higgins, P. Stroeve, and M. P. Srinivasan, *SPIE Proceedings.*, **824**, 121-124 (1988);
b) "Optically Nonlinear Films of Amphiphilic Polymers: Synthesis, Langmuir-Blodgett Deposition, and Optical Measurements," Robert C. Hall, Goeffrey A. Lindsay, Brian Anderson, Stephen T. Kowel, Brian G. Higgins, and Pieter Stroeve, *Mat. Res. Soc. Symp. Proc.*, **109**, 351-356 (1988);
c) "Quadratically Enhanced Second Harmonic Generation in Polymer-Dye Langmuir-Blodgett Films," B. L. Anderson, R. C. Hall, B. G. Higgins, G. A. Lindsay, P. Stroeve, and S. T. Kowel, *Synthetic Metals* , **28**(3), D683-688 (1989);
d) "Second Harmonic Generation in Langmuir-Blodgett Multilayers of Stilbazolium Chloride Polyethers," B. L. Anderson, J. M. Hoover, G. A. Lindsay, B. G. Higgins, P. Stroeve, and S. T. Kowel, *Thin Solid Films* , **179**, 413 (1989).

[13] "NEW HEMICYANINE DYE-SUBSTITUTED POLYETHERS AS NONLINEAR OPTICAL MATERIALS," R. C. Hall and G. A. Lindsay, Navy Case No. 73112 (Oct 1990).

[14] a) "Second-Order Nonlinear Optical Measurements in Guest-Host and Side-Chain Polymer," L. M. Hayden, G. F.Sauter, F. R. Ore, P. L. Pasillas, J. M. Hoover, G. A. Lindsay, and R. A. Henry, *J. Appl. Phys.* **68**(2), 456-465 (1990);
b) "Synthesis and Second-Order Nonlinear Optical Properties of New Coumaromethacrylate Copolymers," R. A. Henry, J. M. Hoover, A. Knoesen, S. T. Kowel, G. A. Lindsay, and M. A. Mortazavi, *Mat. Res. Soc. Symp. Proc.* **173**, 601-606 (1990);
c) "Generation of 315 nm Femtosecond Pulses using a Poled Copolymer Film," D.R. Yankelevich, A. Dienes, A. Knoesen, R.W. Schoenlein, C.V. Shank, and G.A. Lindsay, *CLEO Proceedings*, (postdeadline paper) 602 - 603 (1991).

[15] "FUNCTIONAL AND POLYMERIZABLE COUMARIN DYES FOR COUMARIN DYE-SUBSTITUTED POLYMERS WHICH EXHIBIT NONLINEAR OPTICAL PROPERTIES," G. A. Lindsay, R. A. Henry, and J. M. Hoover, Navy Case No. 72263 (Oct 89).

[16] "NONLINEAR OPTICAL POLYPEPTIDES," G.A. Lindsay, et al., Navy Case No. 73086 (Sept 90).

[17] C.S. Willand and D.J. Williams, *Ber. Bunsenges. Phys. Chem.*, **91**, 1304 (1987).

[18] a) "Nonlinear Optical Polymer with Chromophoric Main Chain," J. D. Stenger-Smith, J. W. Fischer, R. A. Henry, J. M. Hoover, G. A. Lindsay, and L. M. Hayden, *Makromol. Chem., Rapid Commun.* **11**, 141 (1990);
b)"A New Nonlinear Optical Polymer with Mainchain Chromophore," J. D. Stenger-Smith, J. W. Fischer, L. M. Hayden, R. A. Henry, J. M. Hoover, G. A. Lindsay, *A.C.S., Polymer Preprints*, **31**(1), 375 (1990);
c)"Recent Synthetic Developments in Mainchain Chromophoric Nonlinear Optical Polymers," J.D. Stenger-Smith, J.W. Fischer, R.A. Henry, J.M. Hoover and G.A. Lindsay, *Proceedings of the 15th Biennial Polymer Division Symposium* Nov.(1990);
d) "Poly[(4-*N*-ethylene-*N*-ethylamino)-a-cyanocinnamate]: A Non-linear Optical Polymer with a Chromophoric Mainchain. 1. Synthesis and Spectral Characterization," J. D. Stenger-Smith, J. W. Fischer, R. A. Henry, J. M. Hoover, M. P. Nadler, R. A. Nissan, and G. A. Lindsay, *J. Poly. Sci.: Part A: Polym. Chem.*, **29**, 1623-

1631 (1991).

19 "MAIN CHAIN CHROMOPHORIC POLYMERS WITH SECOND ORDER NONLINEAR OPTICAL PROPERTIES," A. Chafin, J. W. Fischer, R. A. Henry, J. M. Hoover, G. A. Lindsay, and J. D. Stenger-Smith, Navy Case No. 72224 (Sep 89).

20 a) "Langmuir-Blodgett Multilayers of Fluorinated , Mainchain Chromophoric, Optically Nonlinear Polymers," J.M. Hoover, R.A. Henry, G.A. Lindsay, C.K. Lowe-Ma, M.P. Nadler, S.M. Nee, M.D. Seltzer, and J. D. Stenger-Smith, *Am. Chem. Soc. Polymer Preprints*, 32(1), 197 (1991);

b)"A New Class of Mainchain Chromophoric Nonlinear Optical Polymers," G.A. Lindsay, R.A. Henry, J.M. Hoover, R.F. Kubin, and J. D. Stenger-Smith, SPIE Proceedings, Dallas, TX, May 8-10 (1991), also reprinted in the Proceedings of the 38th Sagamore Army Materials Research Conference, Watertown, MA, Sept, 10-12 (1991);

c) "Syndioregic Mainchain Polymers for Nonlinear Optical Applications," G.A. Lindsay, J.W. Fischer, R.A. Henry, J.M. Hoover, R.F. Kubin, M.D. Seltzer and J.D. Stenger-Smith, Am. Chem. Soc Polymer Preprints, (Special 40th Anniversary Issue) Philadelphia, PA, June 3-5, (1991);

d) "New Syndioregic, Mainchain, Nonlinear Optical Polymers and their Ellipsometric Characterization," G.A. Lindsay, S.F. Nee, J.M. Hoover, J.D. Stenger-Smith, R.A. Henry, R.F. Kubin, *SPIE Proceedings*, **1064**, 24-26 July (1991).

21 "NEW ACCORDIAN MAIN-CHAIN POLYMERS," G. A. Lindsay, J. W. Fischer, R. A. Henry, J. D. Stenger-Smith, J. M. Hoover, Navy Case No. 72857 (Jun 90).

N92-22477

OPTICAL DATA STORAGE AND METALLIZATION OF POLYMERS

C. M. Roland and M.F. Sonnenschein
Chemistry Division, Code 6120
Naval Research Laboratory
Washington, D.C. 20375-5000

INTRODUCTION

The utilization of polymers as media for optical data storage offers many potential benefits and consequently has been widely explored. The imaging process should produce stable images which have high contrast (readily discernible differences between the image and the background) and high resolution (which allows higher storage density). There are two basic lithographic mechanisms - "photon mode" processes, wherein absorption of a single photon causes an incremental change in the image intensity and "thermal mode", which relies on heat to effect a detectable change in the storage medium. For a photon mode process the spatial resolution is diffraction limited; stray photons falling outside the desired area degrade the resolution and image quality. Hence, photon mode processes must rely on short wavelength radiation (ultraviolet, x-ray, and electron beam) to achieve high resolution. These approaches are employed in much photoresist technology.

An alternative writing mechanism is to use heat to effect physical or chemical changes. Thermal processes selectively raise the temperature to some critical point at which image formation commences. Because of their non-linearity (i.e., the extent of the storage medium's response is not simply proportional to the amount of input energy), thermal lithographic techniques are inherently high in both contrast and stability. Since a single photon provides insufficient energy to induce the marking event, thermal methods are not diffraction limited; consequently, there is no *a priori* reason to employ short wavelength radiation. However, it is anticipated that thermal diffusion away from the directly heated regions may smear the image, and therefore conventional thermal imaging methods are regarded as low resolution techniques.

This report describes new developments [1-6] in thermal imaging wherein high resolution lithography is accomplished without thermal smearing. The emphasis has been on the use of poly(ethylene terephthalate) film, which simultaneously serves as both the substrate and the data storage medium. Both physical and chemical changes can be induced by the application of heat and thereby serve as a mechanism for high resolution optical data storage in polymers. The extension of the technique to obtain high resolution selective metallization of poly(ethylene terephthalate) is also described.

EXPERIMENTAL

All polymers used in this study were obtained from commercial sources. Typical experiments were carried out using polymer sheets produces by extrusion through a slit die; for this reason surface quality and smoothness were mediocre. Heating was accomplished by exposing polymer films (typically 0.005" thick) to 10.6 μm infrared radiation from a Coherent Model 42 CO_2 laser. The output of the laser was at least two orders of magnitude more intense than necessary for the present experiments; hence, substantial attenuation of the beam was required. The intensity profile of the laser beam was nonuniform; thus, reliable estimates of actual power requirements were not possible.

To create an image, the laser beam flood illuminated a mask comprised of gold patterned on a GaAs wafer. The mask lay on top of the polymer, whereby radiation was selectively prevented from impinging on the polymer surface by reflection. In an alternative arrangement, an image was produced by reflecting the laser light from aluminum surfaces onto the films.

For the metallization experiments the film surface was covered with gold, aluminum, or copper prior to irradiation. Both vacuum deposition of the metal, as well as physical lamination of a metal foil against the polymer, were used. When using vacuum deposition, the polymer surface was uniformly covered with non-continuous islands of the metal. This allowed irradiation through the metal-covered layer in order to heat the polymer. For the films covered with a metal foil, the irradiation of course had to be done through the back side.

RESULTS

Laser Induced Crystallization.

Crystalline - amorphous phase transformations are a potential basis for thermal lithographic techniques. The conversion of an amorphous polymer to the crystalline state alters the optical properties, introducing opacity and birefringence, and thus producing an image. A number of studies have employed the crystallization of small molecule species residing on or in a polymeric matrix. The advantage of using a polymer as the active medium is that it can simultaneously serve as the substrate. The polymer film must be initially amorphous, but yet highly crystallizable, in order to produce a high contrast image. Few polymers crystallize extensively above room temperature, while still crystallizing sufficiently slowly to be obtainable in the amorphous state. Among these few poly(ethylene terephthalate) was determined to be unique in its ability to provide high resolution data storage.

Laser induced crystallization was accomplished in PET, for example, by heating the film over roughly 10 μm wide regions with infrared radiation reflected from a piece of aluminum. The localized heating causes the PET to be brought above its crystallization temperature. The crystalline image thus produced appears opaque (white) to the naked eye, while the high birefringence associated with the crystal phase produces a very distinct image when viewed through crossed polarizing filters. The imaging is not accompanied by any alteration of the surface topography of the films. The crystalline images produced by the irradiation were stable to 245°C, at which point melting commences.

In an alternative arrangement a mask was used to define the image. The gold coating comprising the mask image was for convenience placed in physical contact with the polymer film. This arrangement reduces the temperature of the regions shielded from the radiation, since the gold functions as a heat sink. By this method high resolution (*circa* one micron) images, with both good contrast and edge acuity, were routinely obtained. Unfortunately the available masks did not contain images any smaller than this, so that the ultimate limits on the resolution could not be assessed.

The crystalline images produced by the infrared radiation extend completely through the film thickness. There was no evidence of any gradient or accumulation of crystallinity toward the front (irradiated) side of the films. With the mask in contact with the film, higher power levels were necessary to induce crystallization, presumably to compensate for reflection losses at the interfaces (measured to be 30%) and heat conduction to the gold.

The efficiency of the marking process can not be quantified with the present experimental apparatus, primarily because of the non-uniformity of the laser intensity profile. Dose-response curves exhibit the non-linearity expected for a thermal method. This non-linearity contributes to the contrast and edge acuity achievable. Non-linearity of the response is a principle attraction of thermal imaging processes.

The resolution of a thermal process is expected to be governed by the diffusion of heat away from the directly irradiated regions. Modeling of the heat flow in these experiments has been carried out; the results indicate that the material is not in thermal equilibrium as the lithography transpires. This is the reason for the absence of image smearing by thermal diffusion.

An obvious advantage of using a physical change such as crystallization as the writing mechanism is that such a process is reversible, potentially allowing for multiple read/write cycles. This reversibility has in fact been demonstrated. Thermally crystallized images could be erased by reapplication of the laser radiation. The erasure is reversible, with the crystalline image reproduced by exposure again to an appropriate level of the laser radiation. The crystallization and melting can be successively executed by simple adjustment in the intensity or duration of the irradiation.

An initially crystalline, and thus opaque, PET film can be written on by using the laser to locally melt the polymer. A disordering process such as melting can be accomplished significantly faster than the reverse operation of ordering the polymer segments into a crystalline phase; therefore, in principle induced melting of a crystalline film is potentially a faster marking process than crystallization of amorphous polymer.

Ablation

Although lacking reversibility, a chemical change induced by heat can also serve as a lithographic method. When the PET is exposed to higher levels of the infrared radiation, a different process than crystallization occurs. The more intense irradiation promotes significant chemical bond rupture in the PET, with the byproducts of this decomposition expelled as fragments or vapor. This ablation at the PET surface creates a three dimensional image. These images, since they result from a chemical change, are not erasable. The dimensions of the images obtained by ablation were less than 1 μm, with exceptional edge acuity. Although marks slightly less than 1 μm in width were also produced, the masks used to image the laser light have defects at this scale. The ultimate resolution achievable with this technique can not be determined with the presently available equipment.

The depth of the laser etching is proportional to the irradiation time. A correspondence was also found between this depth and the weight loss measured for the films. Similar to the radiation induced crystallization, there is minimal thermal diffusion, as evidenced by the steepness of the walls of the troughs produced by the laser etching. The existence of a threshold for thermal ablation, below which there is no change in the film's appearance, facilitates attainment of high resolution imaging.

The use of thermal ablation as an imaging method was also carried out with several other polymers. High resolution images could be obtained with polycarbonate and poly(ether ether ketone). Using polysulphone, poly(methyl methacrylate), and polychlorotrifluoroethylene, some deformation transpired upon exposure to the laser radiation. Images were achieved in the latter two materials, although an optimal process might involve radiation of a different wavelength. The ablation of two thermosetting polymers, poly(cyanurate) of bisphenol A dicyanante and Epon 28 epoxy was accompanied by some charring of the film. The efficiency and imaging quality varied widely, with the best overall results achieved with PET.

Metallization

It is known that PET will adhere to metals, and particularly to metal oxides, when raised above its melting point while in contact with the metal. Actually, for amorphous PET it is only necessary to heat the polymer above the glass transition temperature for bond formation to transpire. Localized metallization of PET can therefore be effected by its irradiation through a mask. The underside of the PET is maintained in contact with the metal, for example in the form of a thin film or a deposited layer, during exposure to the infrared light. After irradiation the metal in the unexposed regions is readily brushed or stripped away, leaving behind a metallized pattern. Alternately if the deposited layer is discontinuous, the irradiation can be done through the metal layer itself. The bonding to PET is superior for copper and aluminum than for gold. Attempts to remove the former two metals invariably resulted in cohesive failure of the polymer. In all cases the adhesion of the metal to the polymer film passed a standard "Scotch tape test". All indications are that the metallization process is governed by the same factors as the lithography processes described above. Hence, it is expected that the achievable resolution of the film metallization should be better than one micron.

SUMMARY

In anticipation of smearing by thermal diffusion, it is generally believed that photon processes are necessary for high resolution lithography. The present results with infra-red laser irradiation of PET, however, indicate that thermal methods have potential for high resolution optical data storage applications. Both radiation induced crystallization and ablation have been demonstrated to produce high resolution optical images in the polymer film. The two processes prevail at different levels of radiation intensity. While only the former process is reversible, ablation may offer advantages with regard to contrast and resolution. Although the resolution limits of these methods are presently unknown, one micron marking has been routinely demonstrated. These techniques can be readily extended to achieve metallization of polymer films, with the attributes of selectivity and high resolution retained. Particularly noteworthy is the simple one-step nature of the metallization process. No chemical reagents or extraneous washes are necessary.

A familiar commercial application of optical data storage is the compact disk (CD). Information takes the form of a series of pits in the surface of a polymer film, each depression having the same depth but being of different lengths and separations. The variation in the reflection of a light beam as it traverses alternately the smooth CD surface and the pits provides the data transcription. The pits in a CD are created by a stamping process; a master engraves onto the polymer substrate a negative relief. The lithographic methods described herein have the potential to be a "drop-in" replacement for the manufacture of CD's and similar technologies. The non-contacting aspect of the present laser marking techniques, however, will enable one to circumvent the manufacturing limitations of a physical stamping process.

ACKNOWLEDGMENTS

This work was sponsored by the Office of Naval Research.

REFERENCES

1. M.F. Sonnenschein, A.M. Kotliar, and C.M. Roland, Poly. Eng. Sci. 30, 1165 (1990).

2. M.F. Sonnenschein and C.M. Roland, Appl. Phys. Lett. 57, 425 (1990).

3. C.M. Roland, J.P. Armistead and M.F. Sonnenschein in Thermal Marking of Amorphous Poly(ethylene terephthalate); Shalaby, S.; Clough, R., Eds.; ACS Symposium Series; Amer. Chem. Soc.: Washington, D.C., 1991.

4. M.F. Sonnenschein and C.M. Roland, U.S. Patent #4,975,358, 1990.

5. M.F. Sonnenschein and C.M. Roland, U.S. Patent #5,043,251, 1991.

RELATED LITERATURE

1. The Effects of Radiation on High-Technology Polymers, Reichmanis, E.; O'Donnell, J., Eds.; ACS Symposium Series; Amer. Chem. Soc.; Washington, D.C., 1989; Vol. 381.

2. Symposium on Polymers in Microlithography; Poly. Mater. Sci. Eng., 1989, 60, 40.

3. Reichmanis, E.; Thompson, L.F. Chem. Rev. 1989, 89, 1273.

4. Electronic and Photonic Applications of Polymers; Bowden, M.J.; Turner, S.R., Eds.; Advances in Chemistry; Amer. Chem. Soc.: Washington, D.C., 1988; Vol. 218.

5. Symposium on Polymer in Information Storage Technology; Polymer Preprints, 1988, 29, 195.

6. Polymers for High Technology - Electronics and Photonics, Bowden, M.J.; Turner, S.R., Eds.; ACS Symposium Series; Amer. Chem. Soc.; Washington, D.C., 1987; Vol. 346.

7. M.D. Croucher and M.A. Hopper, Chemtech 1987, 17, 426.

N92-22478

LASER DISCRIMINATION BY STIMULATED EMISSION OF A PHOSPHOR

V. K. Mathur
Senior Research Physicist
Naval Surface Warfare Center
Silver Spring, MD 20903-5000

and

K. Chakrabarti
Research Physicist
Naval Surface Warfare Center
Silver Spring, MD 20903-5000
and
Advanced Technology Research Inc.
Laurel, MD

ABSTRACT

A method for discriminating sources of UV, near infrared and far infrared laser radiations has been discovered. This technology is based on the use of a single magnesium sulfide phosphor doubly doped with rare earth ions, which is thermally/optically stimulated to generate colors correlatable to the incident laser radiation. The phosphor, after initial charging by visible light, exhibits green stimulated luminescence when exposed to a near infrared source (Nd: YAG laser). On exposure to far infrared sources (CO_2 laser) the phosphor emission changes to orange color. A UV laser produces both an orange red as well as green color. A device using this phosphor is useful for detecting the lasers and for discriminating between the near infrared, far infrared and UV lasers. The technology is also capable of infrared laser diode beam profiling since the radiation source leaves an imprint on the phosphor that can be photographed. Continued development of the technology offers potential for discrimination between even smaller bandwidths within the infrared spectrum, a possible aid to communication or wavemixing devices that need to rapidly identify and process optical signals.

INTRODUCTION

Availability of large number of infra-red lasers has resulted in the development of very sensitive sensor materials which can detect these lasers. Techniques have been developed to coat these materials uniformly on flat surfaces and to detect the laser either in transmitting or reflecting mode. However, these sensors, though highly efficient in detecting these lasers are unable to discriminate between them. It is desirable in many applications in industry, communication and defense to discriminate, if possible, between the laser wave lengths. Here we present a scheme by which it is possible to discriminate between invisible near infrared, far infrared and UV lasers.

Infrared sensor phosphors are charged by visible or UV radiations prior to the detection of infrared rays. Under infrared illumination, the phosphor emits visible radiation, the upconversion takes place at the expense of charging radiation. Essentially the charging radiation raises the carriers into energetically higher trapping states, from where they are optically stimulated to detrap and recombine with the recombination centers which are generally the impurity ions in the lattice. The observed luminescence is characteristic of these centers. Rare earth ions have been found to be quite efficient recombination centers in these phosphors - the popular phosphors being the alkaline-earth chalcogenides.

MATERIAL PREPARATION

The best amongst alkaline-earth sulfides for laser discrimination, we found through our experiments [1,2], is magnesium sulfide (MgS) doubly doped with rare-earths. Samarium in combination with either Cerium (Ce) or Europium (Eu) or Terbium (Tb) are quite efficient dopants for the discriminator. The infrared detector and discriminator MgS doubly doped with rare earth ions was synthesized in two steps. Initially magnesium sulfate reduced to MgS at 800-850°C by carbon disulfide in the presence of flowing argon. The schematic of the system is shown in Figure 1. When the flowing argon is bubbled through carbon disulfide (CS_2) it carries with it CS_2 which reacts with anhydrous magnesium sulfate to reduce it to magnesium sulfide. The outflowing gases are passed through water which dissolves the sulfur dioxide produced during the reaction. Complete reduction of magnesium sulfate to magnesium sulfide took approximately 2 hours. Magnesium sulfide, thus obtained, is then doubly doped with Sm^{3+} and Ce^{3+} (or Eu^{2+} or Tb^{3+}) by mixing the magnesium sulfide with Sm (in the form of Sm_2O_3) and Ce (in the form of Ce_2O_3) in concentration of 1×10^{-3} and $5x10^{-4}$ mole/mole respectively. The above components were thoroughly mixed by dry grinding in a mortar. Mixture was then fired in alumina boat placed in a quartz tube at a temperature of 950^0C in an inert atmosphere of flowing argon. Argon was passed through concentrated sulfuric acid to remove any water vapors contained in it. Good phosphors were produced after 2 hours of firing.

In another experiment, MgS doubly doped with rare-earth ions was produced in a single firing. Anhydrous magnesium sulfate and the dopants in the form of oxides or chlorides were blended by stirring in a suspension of acetone on a hot plate at 100C until most of the acetone has evaporated. The mixture is then placed on a drying oven at 200C for 1 hour to evaporate the remaining acetone. Next this mixture is placed in alumina combustion boats and fired in a McDanel mullite tube in a tubular furnace at 800^0C in an atmosphere of N_2 or Ar bubbled through carbon disulfide. Complete reduction of sulfate to sulfide takes nearly two hours.

The samples are allowed to furnace cool in an inert atmosphere. The excess sulfur was washed out by immersing the sample in carbon disulfide. Sample is then ground to pass through a standard 200-mesh sieve. Magnesium sulfide doubly doped with rare earth ions was deposited on thin glass slides by the method of sedimentation. A slurry of magnesium sulfide was made in 200 proof absolute alcohol and poured on to well cleaned glass slides. After the particulate matter settled on the slide the excess alcohol was removed from the bath and the films were dried for several hours at 100C. The films were protected from the atmospheric moisture by a coating of Dow Corning 805 binder. These films could be stored for 2-3 years without any degradation.

METHOD OF DISCRIMINATION

The sample is charged by appropriate visible or UV light. This raises the charged carriers into the traps in the material. Infrared radiation stimulates these carriers to detrap and recombine with the charge of opposite sign at the recombination center giving out light characteristics of the center. Periodic recharging of the device is required because the trapped charge leaks away over a period of time on storage or is released by the incident infrared rays. For MgS doubly doped with Eu, Sm, the room light (~500nm) is sufficient to charge it continuously. However for MgS:Ce, Sm near UV light (~320nm) is required to charge it. When near infrared rays impinge on the MgS doubly doped with cerium and samarium, a green emission characteristic of cerium ion is observed. This phosphor responds to 850nm - 1400nm radiations as shown in figure 2. The peak response is around 1060nm. On the other hand 10.6μm can easily heat the phosphor and the resulting emission is orange characteristic of samarium ion (see figure 3). When a UV laser strikes the phosphor, it will emit both cerium characteristic green as well as samarium characteristic orange color. This property of MgS doubly doped with cerium and samarium lends itself to an application in laser discrimination between three groups of lasers lying in near infrared, far infrared and UV regions (see figure 4). By discerning the color of emission through a filter gating system, it is possible to discriminate between invisible near infrared, far infrared and UV lasers. It may be noted that in addition to these emission characteristics, infrared stimulated luminescence will appear as a burst and then decay with time and will require charging or respotting before detecting the subsequent signal. The temporal profile of this emission from the phosphor when subjected to UV irradiation will be similar to the incident UV laser. Moreover, the phosphor would not require any charging and would not decay as long as a steady UV laser is shining on it.

DEVICE SCHEMATIC

On the basis of above description, it is possible to design a laser discriminating device. The schematic of the proposed device is shown in figure 5. The phosphor sensor is continuously or periodically exposed to a charging radiation. When an unspecified laser strikes the phosphor, it stimulates the phosphor emission which is picked up by an optic fiber through a condenser. This optical signal can be directed through either the fiber 1 or fiber 2 by the help of a director. Signal from fiber 1 encounter an appropriate green interference filter, while the signal from fiber 2, an orange filter. Depending upon the nature of the incident laser, the optical signal will reach the photomultiplier (PMT) through either one of the fiber systems or through both of them. Setting on the director will give the information about the color of the emission and therefore, the nature of the laser can be deduced. For further temporal information, the output of the PMT can be coupled to an oscilloscope. If the incident laser is in near infrared further analysis of temporal profile of the emission can lead to subtle distinction in the wavelength of the laser.

EXPLANATION OF THE PHENOMENON

The phenomenon on which the laser discriminator works is as follows: the charging light separates the positive (hole) and negative (electrons) carriers and they are trapped into different sites. For example in Ce and Sm doped magnesium sulfide, after charging, the positive charge (hole) resides at cerium ion and the negative charge (electron) is captured by samarium ion. The near infrared laser optically stimulates electrons from samarium ion, which then, recombine with holes in cerium ion giving out a cerium characteristic green luminescence. On the other hand, far infrared laser such as 10.6μm carbon dioxide laser will heat the phosphor, thereby thermally stimulating the holes from the cerium which eventually recombine at samarium ion resulting in orange samarium emission. Under UV laser both cerium and samarium emissions are observed. These emissions will last as long as the UV laser is striking the phosphor. A model for optically stimulated luminescence (OSL) observed under near infrared laser and thermally stimulated luminescence (TSL) observed under far infrared laser (10.6μm) is shown in Figure 6.

SUMMARY

It is possible to synthesize magnesium sulfide doubly doped with cerium (or euorpium or terbium) and samarium and use it to discriminate between near infrared, far infrared and UV laser by a scheme presented in this paper.

ACKNOWLEDGEMENT

This work was supported by NAVSWC Independent Research funds.

REFERENCES

1. K. Chakrabarti, V. K. Mathur, J. F. Rhodes and R. J. Abbundi, J. Appl. Phys. 64, 1363 (1988)

2. V. K. Mathur and K. Chakrabarti, U. S. Patent #4,947,465 "Method of Laser Discrimination using Stimulated Luminescence", 7 August 1990.

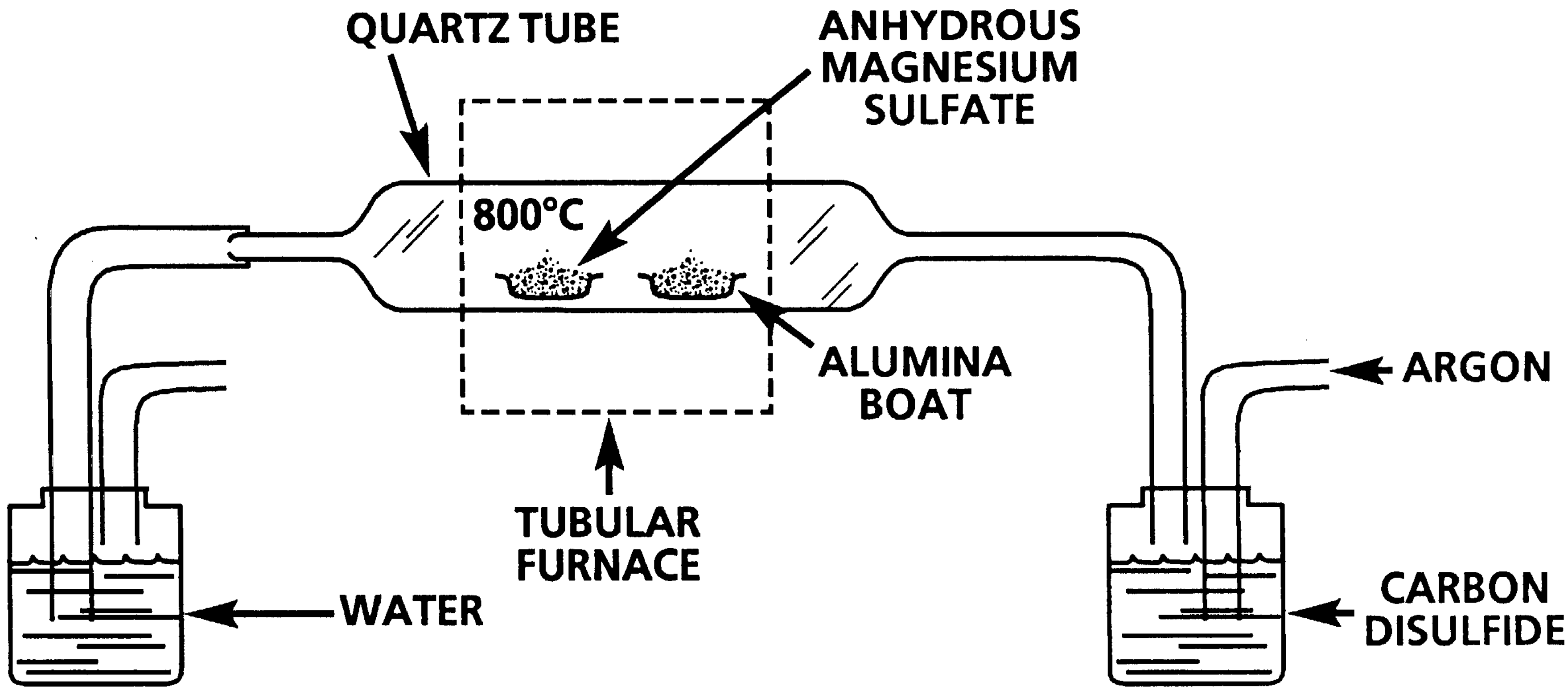


FIGURE 1. SCHEMATIC OF SYNTHESIS ARRANGEMENT

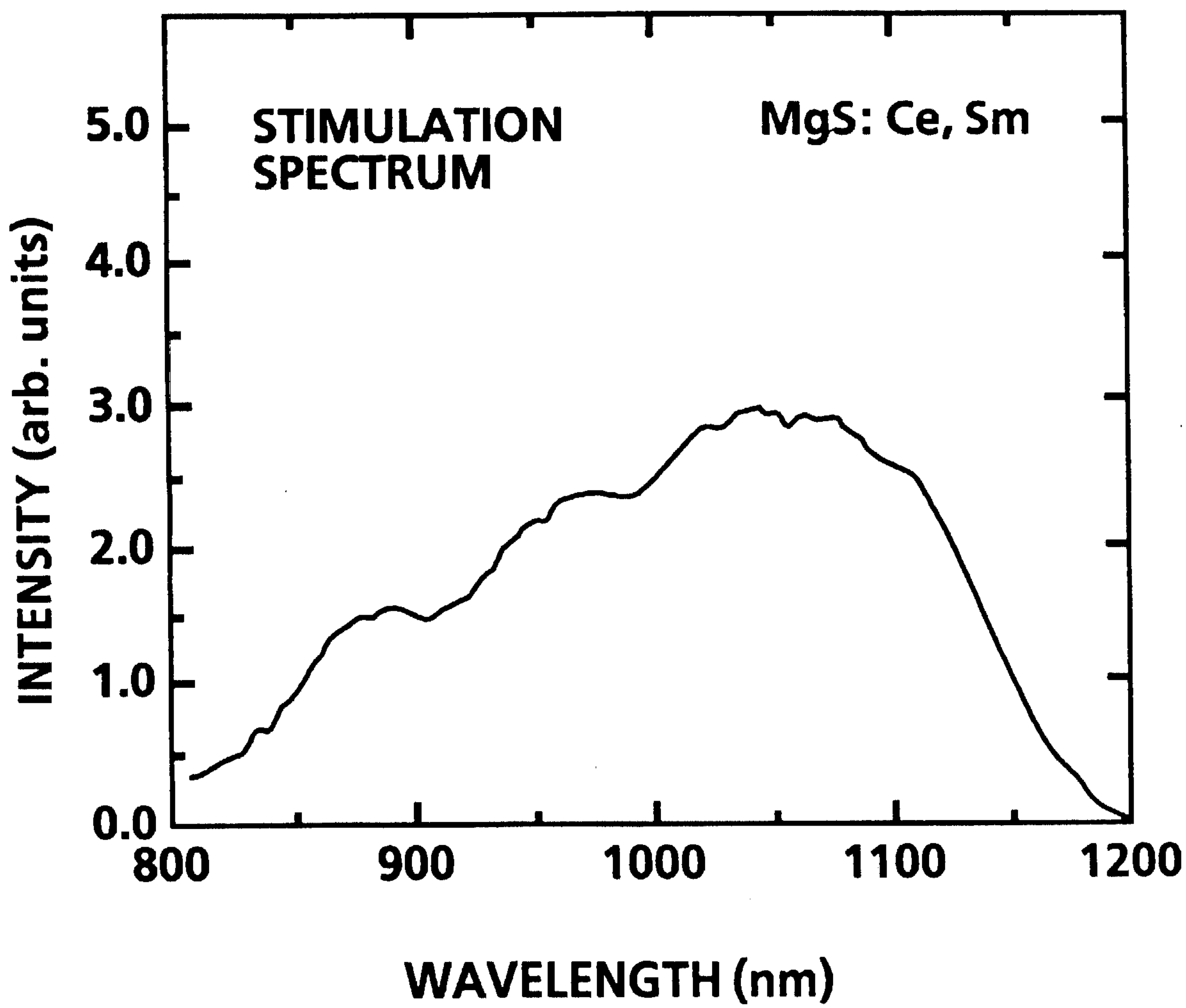


FIGURE 2. STIMULATION SPECTRUM OF OSL EMISSION

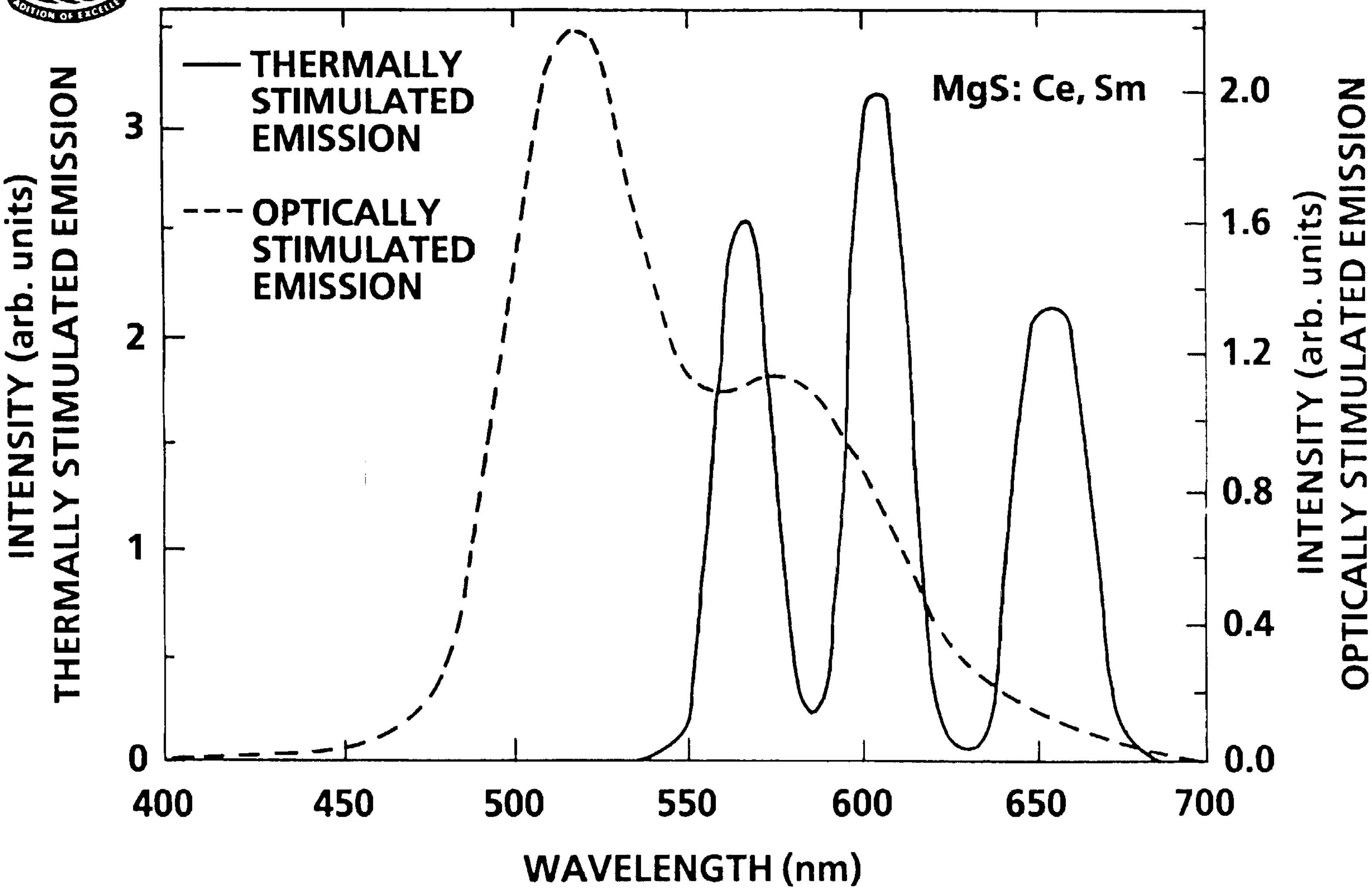

THERMALLY STIMULATED EMISSION
OPTICALLY STIMULATED EMISSION
MgS: Ce, Sm
INTENSITY (arb. units) THERMALLY STIMULATED EMISSION
INTENSITY (arb. units) OPTICALLY STIMULATED EMISSION
WAVELENGTH (nm)
0
1
2
3
0.0
0.4
0.8
1.2
1.6
2.0
400
450
500
550
600
650
700

FIGURE 3.

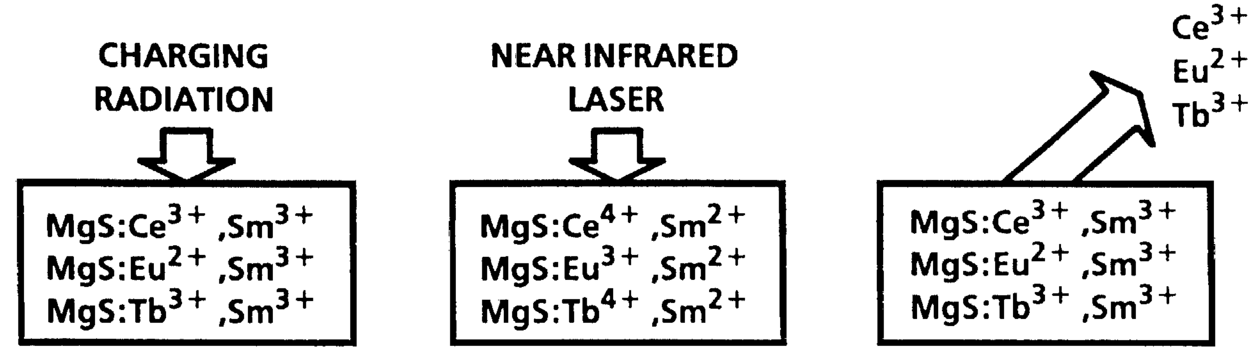


FIG. 4 (a)

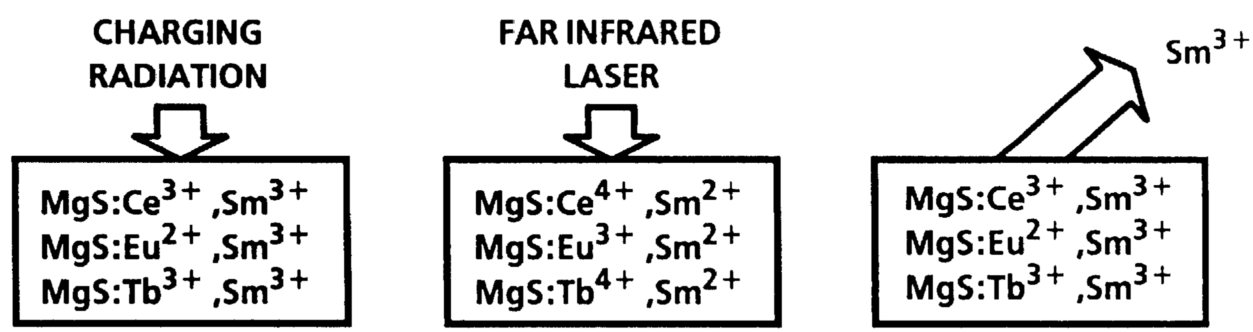


FIG. 4 (b)

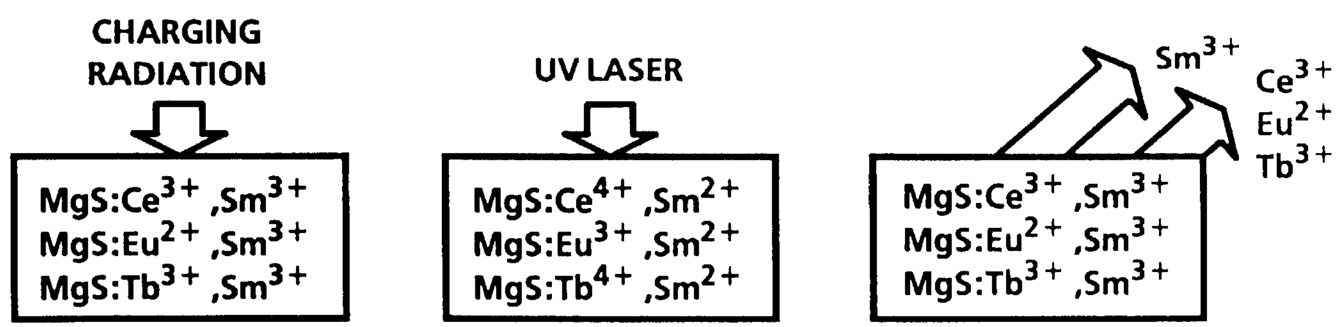


FIG. 4 (c)

FIGURE 4. SCHEMATIC OF LASER DISCRIMINATION

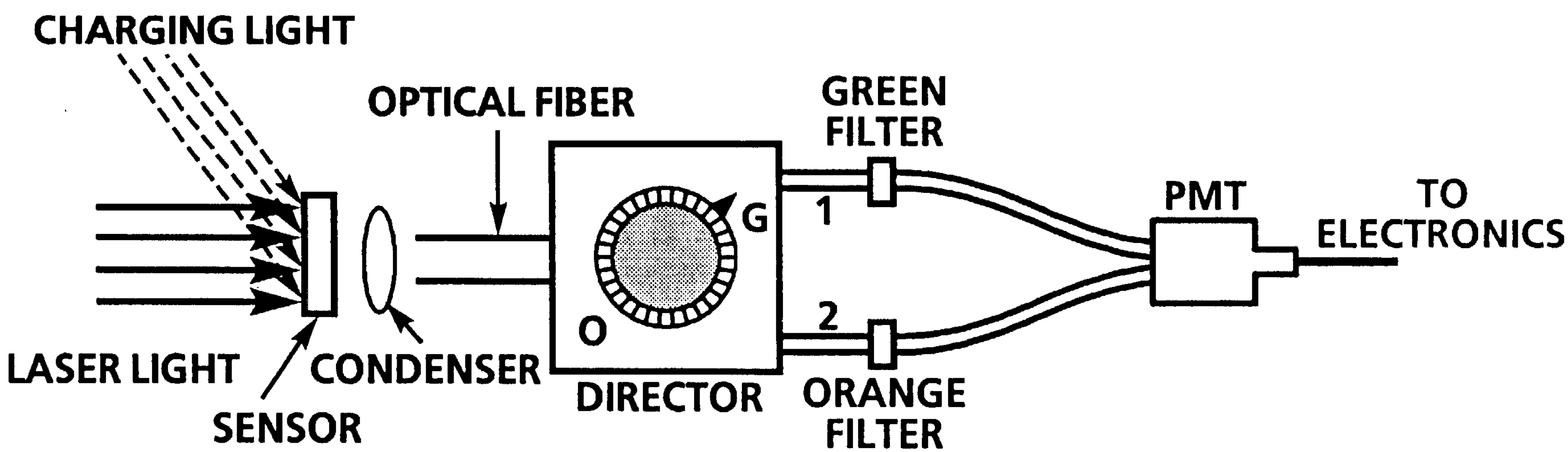


FIGURE 5. SCHEMATIC OF THE PROPOSED DEVICE

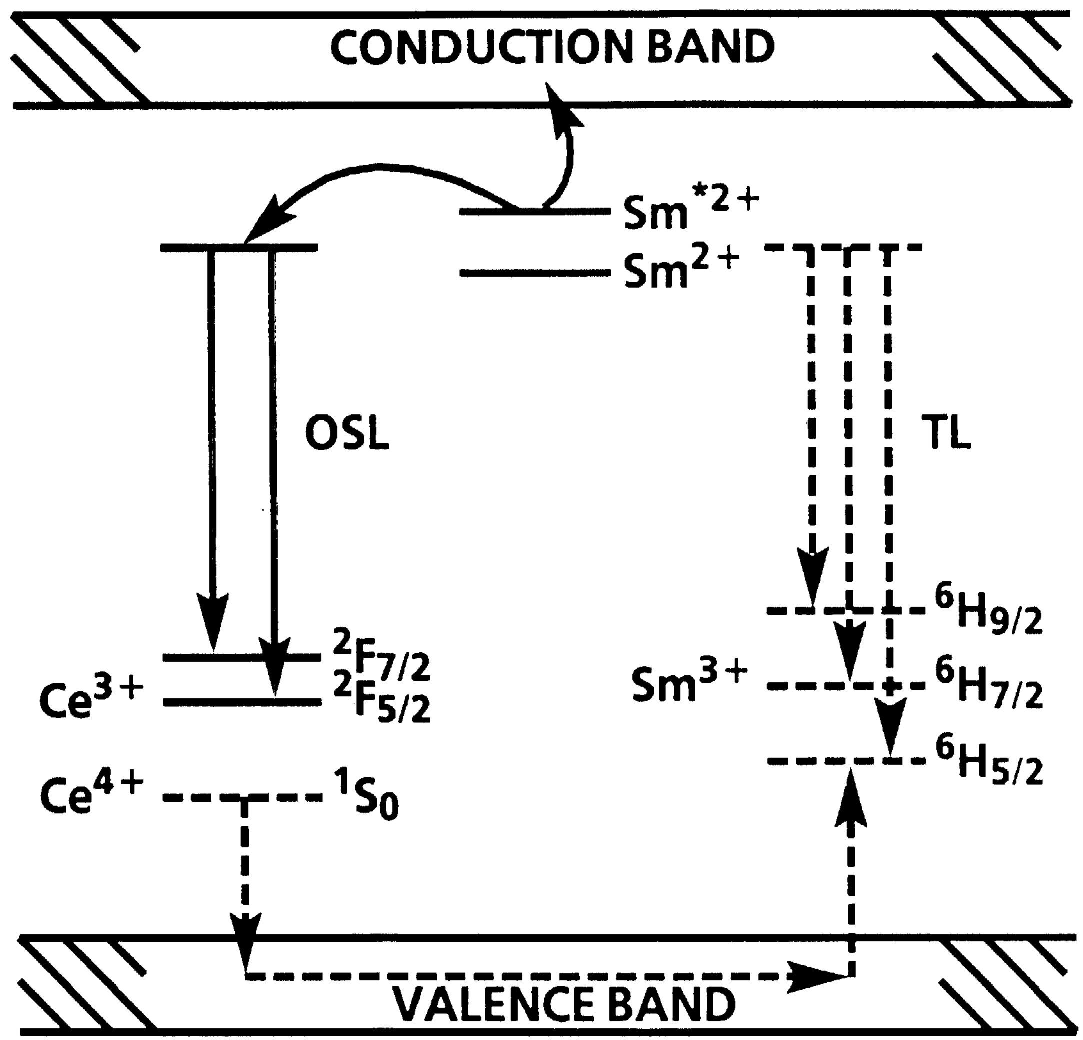


FIGURE 6. A MODEL FOR OBSERVED PHENOMENON

N92-22479

Q-SWITCHED LASER PRELASE DETECTION CIRCUIT

George E. Lockard
NASA Langley Research Center
Hampton, VA 23665

ABSTRACT

A compact electronic circuit has been developed to detect prelasing in Q-switched pulsed laser systems and once detected to shut down the laser before the next laser pulse occurs. The circuit is small, compact and uses a minimum of components which makes it quite economical, thus readily lending itself to commercial applications. It can easily be incorporated into virtually any Q-switched laser system or can be made into a stand alone auxiliary unit. Use of this detection circuit would improve the reliability of a laser system by reducing a source of possible costly optical damage. This paper will discuss the circuit operation and instrument requirements necessary to incorporated the circuit into a laser system.

INTRODUCTION

In Q-switched laser systems there is a condition referred to as prelasing. This condition is one in which there is some amount of uncontrolled laser output occurring along with the normal laser output. Prelasing occurs when laser light "leaks" out of the laser cavity prematurely. This leakage is due to the inability of the Q-switch to completely hold off the lasing action of the laser cavity. Due to the detrimental effects associated with prelasing the condition is considered to be undesirable and needs to be detected and avoided.

The undesirable effects of prelasing fall into two general categories. The first are those directly related to the optical damage effects on the laser system optics and on the optics of the recipient of the laser system output. This optical damage can be quite costly both to the laser manufacturer and the laser user due to both the financial costs associated with the replacement cost of the optics and the system down time necessary to correct the damage. This repair down time is usually quite inconvenient and annoying. The second category includes those problems not related to optical damage effects but to the effects on the applications of a prelasing laser beam. Detrimental effects on the laser output applications would include those associated with multiple output beams, varying energy per laser pulse, changing average output power and output pulsewidth problems. Reliability of a laser system and that of the results derived from the application of the laser output would be improved by using this detection circuit.

CIRCUIT DESCRIPTION

This circuit was specifically designed to detect prelasing in a commercial Nd:YAG pulsed laser operating at a 10 hertz repetition rate and once detected to then shut down the system before the next occurring laser firing. Operation of the circuit is based on the premise that the desired laser output occurs consistently after the Q-switch operation and that any other signal which is present before that moment is to be considered prelasing. All that is needed to integrate the circuit into the laser system is access to the laser light, the Q-switch trigger pulse and the laser security interlock line.

A simplified circuit diagram is shown in figure 1. The basic circuit consists of six sections. They are the optical detection, comparison, delay, sync input, control and output stages. The optical detection stage converts the optical signal into an electrical signal for the comparison stage to trigger off of. The output of the comparison stage is delayed by the delay stage and then sent to the control stage. The Q-switch signal is sampled by the sync input stage which conditions and provides a variable delay to the sync signal which is also sent to the control stage. The control stage compares the timing of the these signals and generates a signal to control the output stage. The output stage interrupts the laser security line and thereby stops the laser.

Optical Detection Stage

The optical detection stage consists of a high speed small area silicon PIN diode detector operated in the reverse biased mode . This mode of operation provides for very fast detector response times and high signal sensitivity. The associated circuitry contains a power supply noise filter capacitor, a bias voltage current limiting resistor and a 50 ohm output termination resistor for the detector . This whole detection stage can easily be located away from the rest of the circuitry so long as proper transmission line principles are adhered to and thereby simplifying the installation of the circuit with the laser system.

The silicon detector used has an optical response from 350 to 1100 nanometers. In flashlamp pumped or frequency doubled systems there are several different frequency optical signals available . In our application we had available the multi-spectrum pulse from the flashlamp, the 1060 nanometer fundamental and the 532 nanometer doubled laser output pulses. A filter can easily be placed in front of the detector to limit the detector signal to just the desired laser light wavelength, which, in our application was the 532 nanometer output pulse.

The circuit will readily work with detectors used for other wavelengths providing they supply voltage signals that are impedance matched to the rest of the circuit. The speed of the detector selected will determine the maximum speed of the system. This enables the basic circuit to be used with practically any pulsed Q-switched laser system regardless of the operating optical output wavelength.

In fact , after the detection stage has converted the optical signal to an electrical signal the prelase circuit becomes totally insensitive to the source of the electronic signal which is sent to the comparison stage. Because of this, the circuit timing could be preset before installation into the laser system.

Comparison Stage

The comparison stage is based around a high speed voltage comparator (IC-1) with TTL complementary outputs. The noninverting input signal is supplied by the detection stage , while the inverting signal is derived from an adjustable voltage divider. The voltage divider provides both positive and negative voltage references for the comparator enabling the comparator to trigger off of either voltage polarity signal from the detection stage. This enables the detection stage to be changed at any time without having to modify the rest of the circuit. The voltage reference was designed around a potentiometer mounted on the front of the circuit enclosure. This provides for operator control of the trigger level after installation of the circuit into the laser system . The external adjustment potentiometer could be mounted on the printed circuit board for preset voltage reference applications thereby saving space and cost.

The comparator integrated circuit has two complementary TTL outputs. The inverted output goes to the delay stage while the noninverted TTL output goes to the comparator monitor . The comparator monitor signal is used by the operator when adjusting the voltage reference potentiometer.

Delay Stage

The next stage is the delay stage which consists of a series of CMOS inverters connected in series . This delay stage generates a fixed delay which is necessary to compensate for the differences in the circuit generated delays between the optical monitor signal and the Q-switch monitor signal. While integrated circuit nanosecond delay lines are available for this purpose , the inverters were selected due to their availability and low cost over them. Another advantage of the use of the inverters is the pulsewidth stretching effect inherent with their usage. Although this increase in pulsewidth limits the minimum detection time between the prelase signal and the lasing signal, it consistently ensures adequate pulsewidths for the control stage signals especially when dealing with extremely narrow laser pulses.

At this point the optical signals have been detected , converted to digital voltage levels and delayed. These signals go to the control stage along with the modified Q-switch monitor signal from the sync stage.

Sync Stage

The sync stage consists of two digital logic circuits. The first is a high speed CMOS Schmitt-Trigger inverter with pull-up resistor that is used to sample the Q-switch trigger signal while buffering the circuit from the laser electronics. It's output goes to a CMOS dual non-retriggerable monostable multivibrator integrated circuit (IC-5) whose first multivibrator section is connected to trigger off the inverter's output signal's falling edge . The output pulsewidth of the first multivibrator is controlled by the R/C time constant of a fixed capacitor (C1), fixed resistor (R1) and variable resistor network . The fixed resistor sets the minimum pulsewidth while the variable resistor is used to vary the pulsewidth from that minimum .

The variable pulsewidth output pulse from the first multivibrator is used to trigger the second multivibrator whose output pulsewidth is set by a fixed R/C network (R2,C2). The second multivibrator is designed to trigger off of the trailing edge of the output pulse of the first stage which provides a variable delay from the original Q-switch signal that is equivalent to the variable pulsewidth of the first multivibrator . This fixed pulsewidth output pulse from the second multivibrator goes to the control stage as the Q-switch sync signal.

In this configuration the dual multivibrator chip serves two functions . It provides a trigger signal to the control stage whose pulsewidth is independent of the original Q-switch pulsewidth . Also it provides a variable delay for the trigger signal . It is this variable timing adjustment that is used to set the optimum timing for the control stage.

Now that the Q-switch monitor sync signal has been generated , it along with the previously described optical monitor signal are integrated in the control stage.

Control Stage

The control stage is made up of a CMOS dual J/K flip-flop (IC-4) with preset and clear options. The first flip-flop is connected to toggle off the optical monitor signal with the sync signal acting as a clear control signal. The timing of the sync signal is set just slightly before the normal lasing pulse. Refer to the timing diagram in figure 2. The sync signal holds the clear line of the flip-flop high and as long as the clear line is held high the flip-flop ignores the toggle input from the optical monitor signal. But should a toggle pulse occur before the Q-switch monitor pulse can clear the flip-flop then that signal will toggle the flip-flop generating a square wave output signal pulse. This earlier pulse would be the pulse generated due to prelasing.

During normal lasing the first flip-flop generates no output signals while during prelasing it generates a pulse whose pulsewidth varies as a function of the time between the prelase pulse and the Q-switch sync monitor.

This output pulse acts as a toggle signal for the second flip-flop. Like the first, the second flip-flop is also connected in the toggle mode except that once toggled the flip-flop must be manually reset before its output will change . The clear command of the second flip-flop is controlled from a three position switch (on-off-momentary) whose functions are run , by-pass and reset.

In the run mode the clear function is held inactive so that the flip-flop output toggles normally on command . The by-pass mode holds the clear function activated so that the output of the flip-flop never changes state regardless of its input . The reset mode provides a manual clear command to reset the output of the flip-flop after being toggled in the run mode .

A resistor (R3) and capacitor (C3) make up a R/C network which is connected to the clear command of the second flip-flop to provide a start up delay for the clear line. This ensures that the flip-flop is always in the deactivated mode after being powered up.

A trigger monitor signal and a triggered indicator are provided by an inverter and LED respectively. The monitor signal from the inverter is used by the operator in setting up the timing relationship of the signal monitor and Q-switch sync signals to the flip-flops. During normal operation, the LED provides an indication to the laser operator why the laser has shut down.

Output Stage

The output of the second flip-flop goes to the output stage which consists of a DIP relay (IC-7) which opens up the laser security line when energized and thereby shutting down the laser.

SETUP AND ALIGNMENT

The degree of circuit sensitivity is dependent on the trigger level and placement of the detector . The detector sensitivity is extremely high and can easily become saturated resulting in distorted output pulses. Care must be taken in the placement of the detector to avoid this.

Alignment of the detector is not very difficult and is done with the control switch in the bypass mode and the laser operating normally. The output of the detector , which is connected to the circuit, is monitored on a scope during its positioning. The operator ensures that the detector is not saturated and that there is ample signal for the circuit. Another channel of the oscilliscope is then connected to the comparator monitor. The trigger level is adjusted until a signal is detected from the comparator monitor.

Next the circuit timing is adjusted. The timing monitor signal is observed on an oscilloscope which is triggered on the laser Q-switch sync signal. The delay potentiometer is adjusted so that a square wave pulse is observed on the oscilloscope. When the laser is not prelasing an output pulse on the timing monitor indicates that the variable Q-switch sync signal is resetting the control stage after the normal lasing pulse occurs. The operator simply adjusts the delay potentiometer. As the timing approaches the desired position the pulsewidth of the timing monitor signal will reduce. At the point where the timing monitor pulse disappears the correct timing relationship between the control stage signals has been achieved . The flip-flop is being reset by the sync signal pulse before the lasing pulse can trigger the circuit. The operator switches the control switch to the run mode and removes the oscilloscope connections.

PERFORMANCE SUMMARY

This prelasing detection circuit has been tested on a commercial Nd:Yag whose output was frequency doubled to 532 nanometers and also on a Nd:YLF research laser . Testing to determine the timing relationship parameters was done electrically in order to prevent possible optical component damage to the laser system . This was done by injecting electrical signals from a signal generator into the optical signal input to simulate the lasing and prelasing signals from the detection stage. The prototype circuit was able to reliably trigger off of input signals to the comparator that were 10 nanoseconds wide and separated by 80 nanoseconds. This separation time was longer than expected from the integrated circuit data sheets and is attributed to inadequate circuit board construction techniques. A circuit board designed with proper high frequency considerations should improve the circuit response and result in faster trigger times (shorter pulse separation). Suggestions for improved speed would include conversion to faster digital integrated circuit logic, the use of nanosecond delay lines and better circuit board construction techniques than those used for the prototype. Minimum separation times between the lasing and prelasing pulses of 20-25 nanoseconds would be expected.

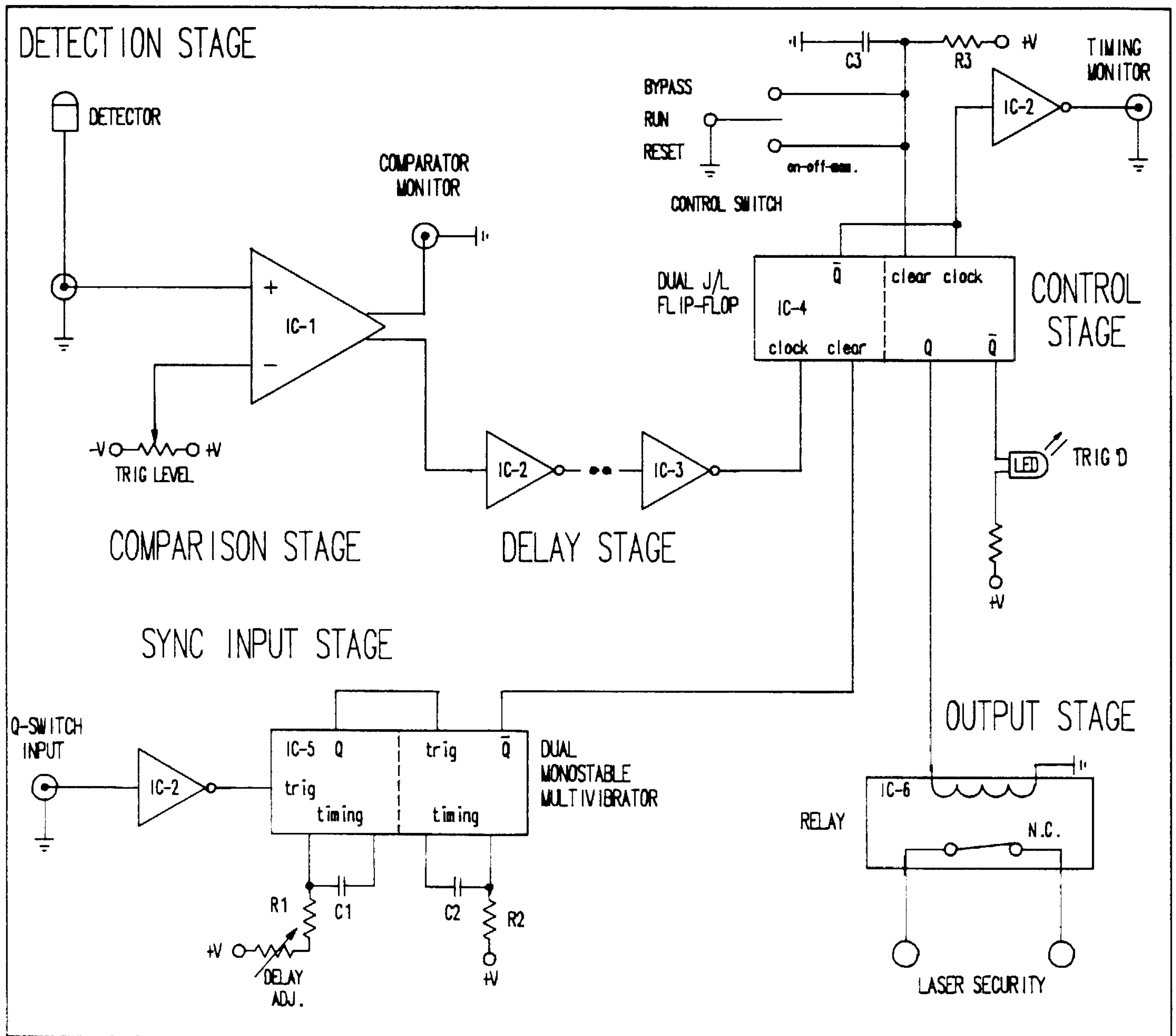


Figure 1: Simplified Circuit Diagram

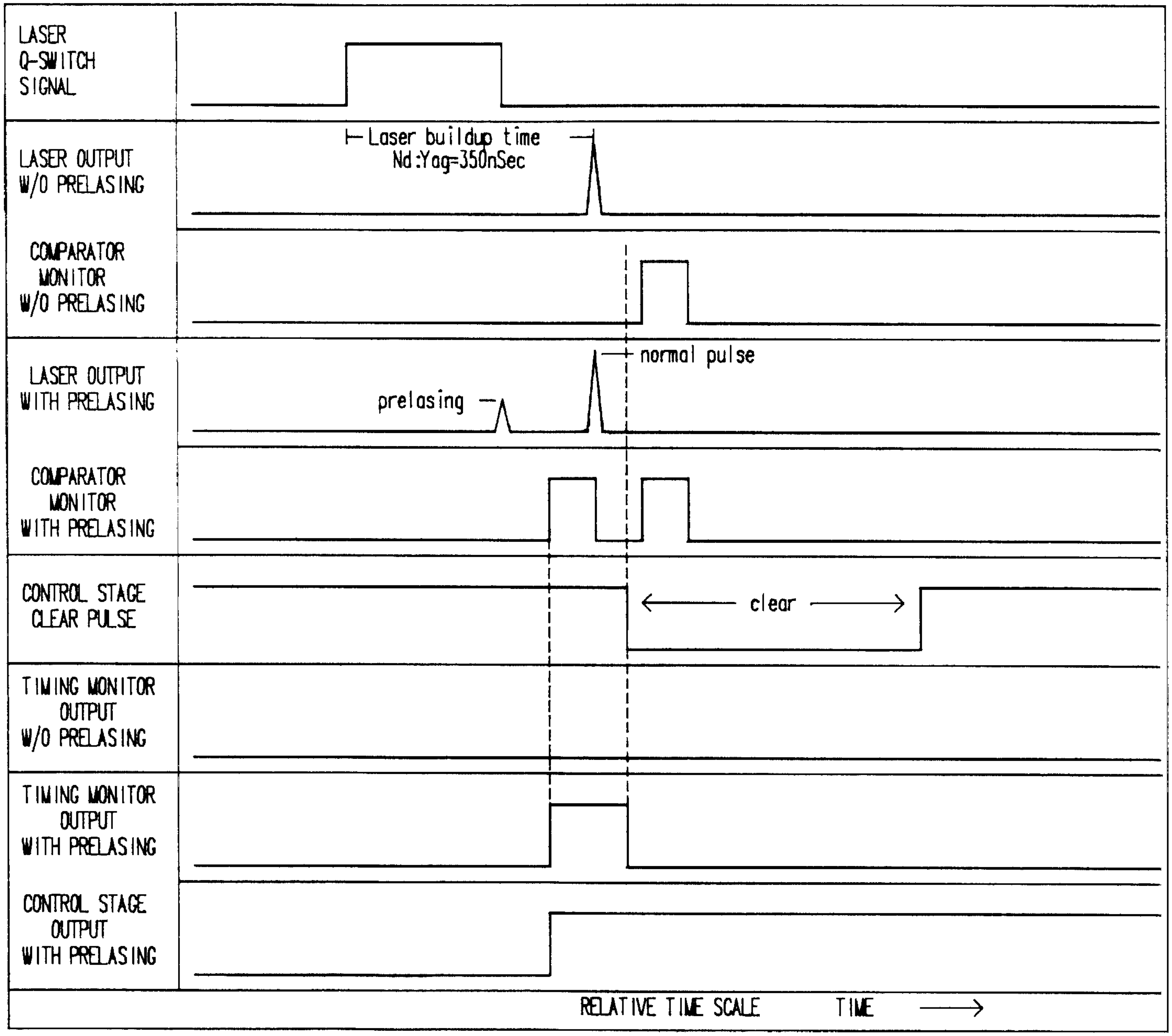


Figure 2: Timing Diagram

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

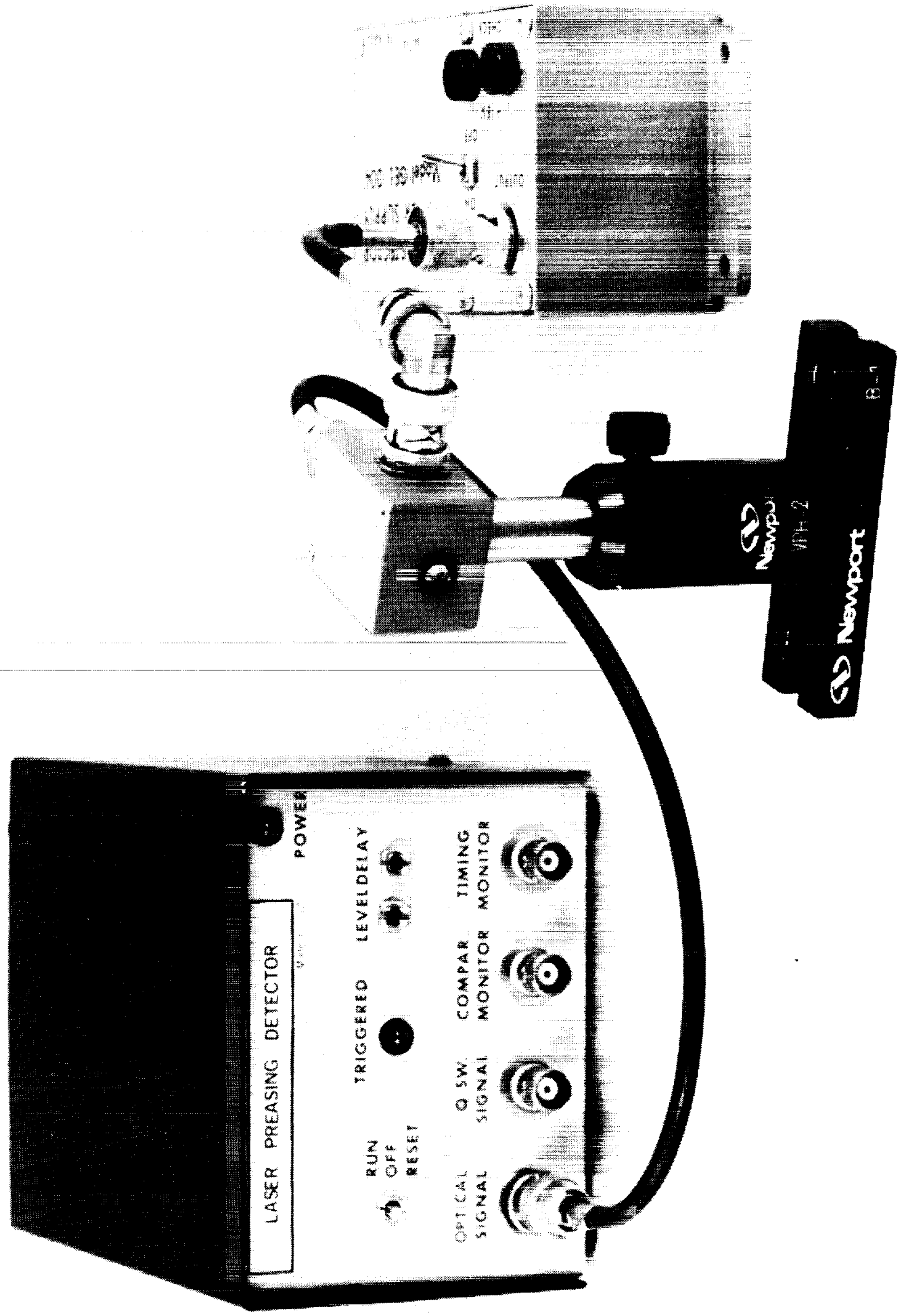


Figure 3: Photo of Detector Instrument and Optical Detector

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

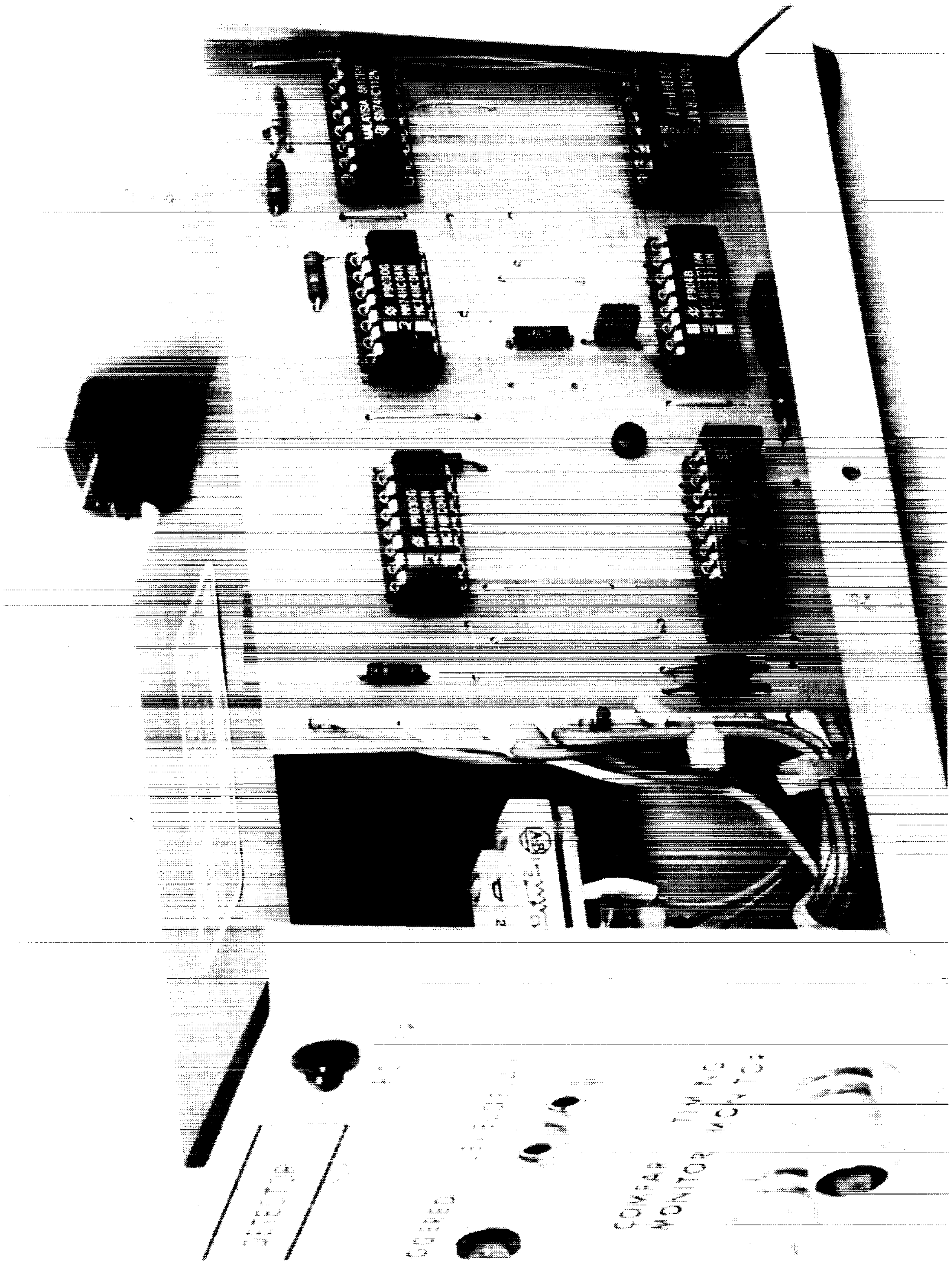


Figure 4: Photo of Close-up of Circuit Board

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

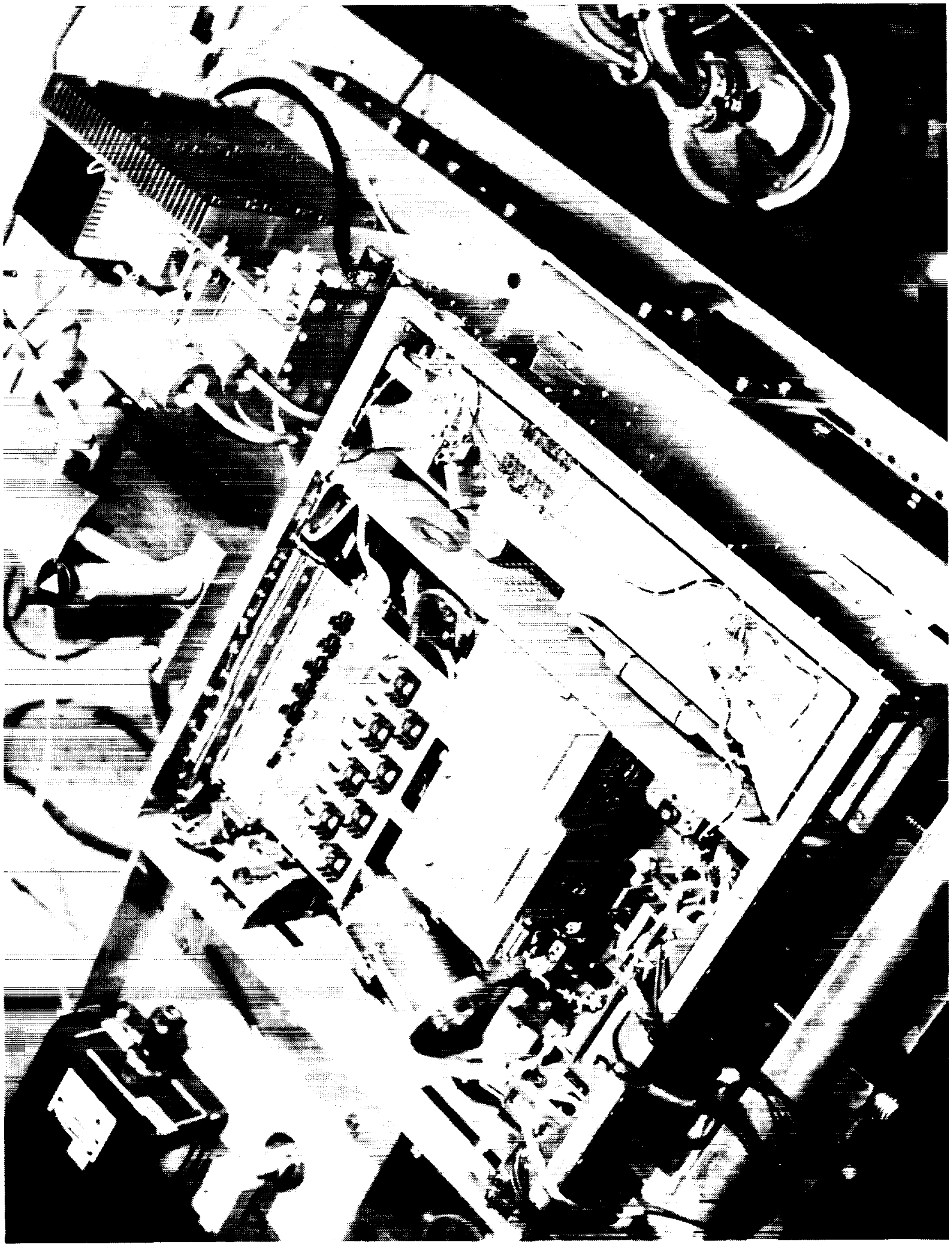


Figure 5: Q-Switched Laser Prelase Detection Circuit

LIFE SCIENCES

(Session C3/Room C1)

Wednesday December 4, 1991

- Application of CELSS Technology to Controlled Environment Agriculture
- Advanced Forms of Spectrometry for Space and Commercial Application
- Ion-Selective Electrode for Ionic Calcium Measurements
- A 99% Pure Molecular Sieve Oxygen Generator

N92-22480

APPLICATIONS OF CELSS TECHNOLOGY TO CONTROLLED ENVIRONMENT AGRICULTURE

Maynard E. Bates, PhD
The Bionetics Corporation
NASA/Ames Research Center
Mail Stop 239-6
Moffett Field, CA, 94035

David L. Bubenheim, PhD
NASA
Ames Research Center
Mail Stop 239-11
Moffett Field, CA, 94035

ABSTRACT

Controlled Environment Agriculture (CEA) is defined as the use of environmental manipulation for the commercial production of organisms, whether plants or animals. While many of the technologies necessary for commercial success of CEA are still in the early stages of development, investment in greenhouses and aquaculture systems in North America is nevertheless doubling approximately every five years. Economic, cultural, and environmental pressures all favor CEA over field production for many non-commodity agricultural crops. Many countries around the world are already dependent on CEA for much of their fresh food (e.g. Holland, Israel, Saudi Arabia, Japan).

Controlled Ecological Life Support Systems (CELSS), under development at NASA's Ames Research Center, Kennedy Space Center, and Johnson Space Center expand the concept of CEA to the extent that all human requirements for food, oxygen, and water will be provided regenerated by processing of waste streams to supply plant inputs. The CELSS will likely contain plants, humans, possibly other animals, microorganisms and physicaly and chemical processors In effect NASA will create engineered ecosystems.

In the process of developing the technology for CELSS, NASA will develop information and technology which will be applied to improving the efficiency, reliability, and cost effectiveness of CEA, improving its resources recycling capabilities, and lessening its environmental impact to negligible levels.

NASA - ADVANCED LIFE SUPPORT

Controlled Ecological Life Support System (CELSS)

A Controlled Ecological Life Support System (CELSS) is a regenerative system which incorporates biological, physical and chemical processes to support humans in extraterrestrial environments. The key biological processes in such a system are photosynthesis and transpiration. Green plants utilizes light energy (radiation) to produce liquid quality water, food, and oxygen while removing carbon dioxide, water, and inorganic elements for air and waste streams. Figure 1 is a diagram of the CELSS concept.

The CELSS concept involves resource recycling by regenerative systems. In an ideal CELSS all materials are cycled between crew and regenerative processors: no new mass would need to be added to the closed system. The only essential input is energy. The CELSS system results from the management of the fluxes of mass and energy while the total mass of the system remains constant.

Development of a CELSS requires identification of the critical requirements that will allow the system to operate with stability and efficiency. When this information is incorporated into electronic expert systems the whole CELSS can be operated as a life support machine, responsive to the needs and activities of the human crew. A major objective or the CELSS R&D efforts at NASA/Ames is to characterize the ability of a plant-based system to provide food, oxygen, purified water, and carbon dioxide removal from closed environment for the purpose of life support. The critical requirements of the plant-based system are genetic, environmental, and cultural. They are the conditions of light, air, and the nutrient solution which control plant growth. They are bred into the gene pool of the selected cultivar. They are the specific processes of crop management known in the past as farming.

Plant Growth Chambers

The primary tool of the plant growth researcher is the growth chamber. This is essentially a container holding a volume of air at controlled conditions of temperature, carbon dioxide concentration, and water content (humidity), a source of photosynthetically active radiation (light), and a container in which roots are bathed in nutrient solution. The light source is a combination of several electric lamps of various types. The nutrient solution consists of water, oxygen and salts of thirteen or more elements essential to plant nutrition (hydroponics). All systems are managed by controllers to maintain the desired plant growth environment. Plant growth chambers allow us to observe and measure the response of specific populations of plants to environment and cultural practice. Such systems are "leaky" - while conditions are maintained, mass flows into and out of the system, sometimes at rapid rates. Therefore the surrounding environment acts as a sink for evolved oxygen, various hydrocarbons, and transpired water. Sources of water, carbon dioxide and elemental ions must be provided continuously.

Crop Growth Research Chamber (CGRC)

In order to study the response of plants to conditions existing in a CELSS it is necessary to build a growth chamber which is closed with respect to mass; That is, no exchange of mass with the ambient environment surrounding the chamber is allowed. Gases must be carefully added and removed and internal pressures controlled.

NASA/Ames Research Center is currently constructing one of the first such chambers in the world (Figure 2). Named the Crop Growth Research Chamber it will give researchers total control of the plant environment within a volume closed with respect to mass. The CGRC is for the study of plant growth and development under stringently controlled environments isolated from the external environment and is designed for the growth of a community of crop plants. The CGRC is the individual unit where various combinations of environmental factors can be selected ant the influence on biomass, food and water production and oxygen/carbon dioxide exchange of crop plants investigated. Sever Crop Growth Research Chambers and laboratory support equipment provide the core of a closed systems plant research facility. This facility will be utilized for research, volatile gas analysis and trace gas challenge to the plants , technical studies (development and evaluation of technology), system control system modeling (development and validation), and system operation.

The advantage of the closed growth chamber is the ability through sensors and analyzers to achieve immediate feedback on the status of plant growth; That is, we can monitor crop growth and its response to environmental manipulations in real time. Having generated a response surface of crop growth to the environment we can then build and test models for management of the crop as a production system. Furthermore, the models and sensors can be incorporated into expert control systems which continually adjust environment for stage of growth, cultural practices, system perturbations, etc. Finally, a human crew can be simulated and the CGRC run as a life support system in response to the simulation. This results in a preliminary test of a CELSS-type system

CONTROLLED ENVIRONMENT AGRICULTURE

Definition

Controlled Environment Agriculture (CEA) is the commercial production of a population of genetically uniform organisms through control of the growth environment. Many kinds of organisms have been grown in CEA including: microorganisms, mushrooms, plants, fish, and other animals. We are specifically interested in food producing crop plants. By using the information gathered from growth chambers to optimize the growth of plants we have been able to generate extremely high crop yields. Table 1 shows some of the yields which have been achieved in CEA systems in comparison to field production.

Growing Organisms as a Manufacturing Process

The primary objective of CEA is to transform the production of crops into a manufacturing process having consistent and predictable output volume , cost of production, and product quality (Figure 3). As in all manufacturing systems quality and consistency of product impact on sale price, which in turn affects Return on Investment. Production efficiencies, reflected as higher yield or reduced cost to produce, also directly impact Return on Investment. Therefore CEA, unlike traditional farming, carries a lower risk of crop failure and a high degree of assurance of product quality and market timing.

In addition to economic and market considerations there are other reasons to use Controlled Environment Agriculture for crop production:

Production in Harsh or Inconsistent Environments

Good arable land with a consistent climate year around is almost nonexistent in this world. For this reason crops are only grown at certain locations in specific seasons, usually well below optimum conditions. Fresh products go in and out of season and harvest areas migrate geographically with the seasons. Deserts, cold climates, and extremely cloudy areas produce little or nothing. Droughts, storms, heat, frost, insects, and diseases kill crops and damage products. Salt buildup in some irrigated soils makes cropping impossible.

CEA cropping allows maximum productivities to be achieved on the Arabian Peninsula, in the US Desert Southwest, in Alaska, Canada, Northern Europe, and Australia. Almost anywhere in the world. Fresh products can be grown close to markets instead of being transported thousands of miles. Persons in isolated environments, e.g. Antarctica, can grow what they need. Persons in environmentally disadvantaged areas can become agricultural producers. All that is needed is a site, a well-designed CEA system, water supply adequate for plant transpiration, and a source of energy.

Land Use Competition

The world is becoming more crowded. Businesses and houses, towns and cities are encroaching on agricultural land, particularly where climates are best. In the United States large areas of agricultural land in California, Florida, and other states are being developed. Desertification in some parts of the world (Middle East, Africa) is removing land from production. Population growth requires more and more production on less and less cultivated land (China, Japan, Europe, Russia, United States). It is obvious that more intensive cropping is going to be required. Some of the world's most populous countries are already heavily committed to the development of CEA (China, Japan).

Resources Management

We have already cited arable land as a diminishing resource. CEA is conservative of land, producing up to 20 times as much product annually as traditionally cultivated land and able to equal or better that production on land where nothing would grow otherwise.

As mentioned previously, the only water needed to grow crops in CEA is that which is transpired by the plants. There is no waste due to runoff or evaporation of irrigation water. Furthermore, use of carbon dioxide supplements to enhance plant growth results in substantial reductions in water use.

CEA fertilizers consist of water soluble salts which are continually circulated to the plant roots. None are lost to runoff. The only use is that which is incorporated into plant tissue.

The one resource which CEA systems appear to require in substantial quantities is energy. In greenhouses this energy is used for heating and cooling(ventilation) to maintain growth environments. Energy use appears high because it is concentrated into a small area, when figured on the basis of use per pound of food produced, energy use in controlled environment agriculture systems is not excessive and is clearly economically sound. However, this too can be conserved. Waste heat from other manufacturing systems, e.g. Archer Daniels Midland's corn sweetener plant in Illinois, or from electrical power plants, e.g. Pennsylvania Power and Light's Montour station in Pennsylvania, can be recovered to heat greenhouses. Off peak energy, excess power generated during certain times of the day, can also be effectively used in CEA especially where electric lighting is used, since supplemental lighting can be done at night. Solar heat can be used both for heating at night and for daytime electricity. Burning of sawdust and wood chips, waste products from forest industries, is used to power greenhouses in Minnesota.

Pollution Control

Closed environments can effectively exclude pests and diseases they may carry. Control of environment provides a means of controlling insects and diseases. Highly oxygenated, healthy root systems and well grown plants are more resistant to diseases. Elimination of soil removes a major source and incubator of pests and diseases. Efficient sterilization of nutrient solutions, as with UV light, can be accomplished in hydroponics. Natural pest control methods, such as predatory insects, work very well in greenhouses. All of the above result in the virtual elimination of the need for chemical pesticides and thus also eliminate potential for pollution of growing systems and the general environment.

Plant nutrients in CEA are water soluble salts composed of ions of thirteen elements: nitrogen, phosphorus, potassium, calcium, magnesium, sulfur, iron, boron, copper, zinc, manganese, molybdenum, and chlorine. All are essential to plant growth. Sometimes other elements such as silicon, aluminum, sodium, cobalt, chromium, etc. are added as trace elements. Of all these ions the ones which cause the most problems when released to the environment are nitrogen (as nitrate or ammonium) and phosphorus (as phosphate) because they constitute the bulk of the salts used. Runoff from fertilized fields and leaching into water tables has been a serious concern for both farms and CEA facilities. Recirculating hydroponics used in a closed system or a well-designed CEA system does not have excess nutrient ions dumped out into the environment. Salts are added at the rate which plants extract them from solution and incorporate them into biomass.

In an ideal CELSS all solid wastes are processed back into carbon dioxide, water, and nutrient salts. Most of this can be accomplished through low technology processes such as drying (evaporation) and burning (combustion). The only problem with combustion is that the nitrogen mostly ends up in gaseous forms and some of the resulting ash is not water soluble. In CEA solid wastes have become a problem. Growing media, old plants, roots, and unsold products must be removed and disposed of. Efficient methods of composting or drying/combustion are definitely needed.

Greenhouses and Plant Factories

Greenhouses are those CEA facilities which use sunlight as the primary source of photosynthetic photon flux for plant growth. They consist of some transparent or translucent material (glazing) fixed into a framework. Glazing is usually glass or plastic. Many different configurations and levels of technology are being used. The advantages gained are protection from the extremes of the ambient environment, provision of a contained volume of atmosphere which can be controlled to some advantage, such as by elevation of temperature and carbon dioxide concentration, resulting in increased efficiency of use of available sunlight. Crop yield, quality, consistency, and predictability are improve as a result.

All greenhouses are driven by the availability of sunlight and require heating and cooling to maintain a good plant growth environment. When available sunlight is insufficient - winter, cloudy days - electric supplemental lighting may be employed to provide additional growth. The design and use of lighting systems has become a major area of research and development in greenhouse crop production.

Some plant growing entrepreneurs have attempted to avoid the vagaries of sunlight and environment by completely enclosing their growing systems so that total control of the environment, including light, is achieved. These "plant factories", essentially large versions of plant growth chambers, have taken a major step toward making crop production a manufacturing process. In doing so they are forced to face questions of lighting, environmental control, plant growth, and economics. Only a few commercial units are currently operating in the world and most of those grow only bedding plants.

The World View

The present and future of Controlled Environment Agriculture in the World is excellent. Holland has about 25,000 ha of greenhouses, 11,000 ha of which are in vegetables. Israel and Saudi Arabia produce their own supplies of fresh vegetables, nearly all from greenhouses, and export to other countries. Columbia exports huge quantities of fresh flowers grown in greenhouses, mostly to the United States. Japan and China are moving rapidly into greenhouse production. Canada, having no warm climates for winter vegetable production, has developed a successful greenhouse industry both in vegetables and flowers. In the United States the greenhouse area is doubling about every five years and is now the fastest growing segment of American agriculture.

A large part of the technology for this growth comes from Europe, especially Holland and Denmark. However, it is a young industry and its commercial development, and that of the more sophisticated plant factories, will require a great deal more understanding of the control of environment, the growth of plants, and the engineering and manufacture of efficient systems for both.

CONTRIBUTIONS OF NASA CELSS RESEARCH AND DEVELOPMENT

Exploring the Limits of Plant Productivity

In our quest to determine the potential of plant crops in life support we will push them to their productive limits. Defining those limits and the environmental parameters with which they are associated are keys to achieving crop production potential.

Environmental Response Surfaces for Specific Cultivars

We already know that each genetically different plant population has its own characteristic responses to environment. It is necessary to learn how that particular population will interact with environment if it is to be incorporated in a life support system. This set or responses is called the environmental response surface for the crop. The methods for determining response surfaces for specific plant cultivars and the descriptions of such surfaces are valuable elements in improving the operating efficiency of CEA.

Models for Management of Controlled Environment Cropping Systems.

Response surfaces and other information about the growing system and plant environment are incorporated into models which allow both predictability and control of crop growth. Development and testing of such models will be a major CELSS activity. The application of the resulting models to CEA will be a substantial new business.

Methods of Mass and Energy Management.

In addition to plant growth models other models will be required to describe the movement of mass and energy through a CELSS. The availability of both is going to be severely limited in space and efficiency of management is essential. As part of the CELSS development we will be learning how to monitor and efficiently control mass and energy fluxes. The resulting models, methods, and systems should be very useful in building better plant factories.

Plant Growth Lighting

One of the energy-based systems mentioned above is plant growth lighting. In a CELSS we need to get the most out of the energy input to plant lighting. This will require both improved engineering of lighting hardware and improved utilization of the light delivered to plants. As an example, light utilization efficiency has been increased by four times in many crops. The lighting systems, cultural methods, and models which describe their interaction will all be of great use to CEA.

Specific Ion Monitoring

In the area of mass balance there is a specific need for sensors/analyzers which can reliably monitor the concentrations of nutrient ions in solution over extended periods of time. In order to maintain plant growth rates the essential ions must be balanced. None should be excessive and potentially toxic or deficient. Differential rates of uptake can upset this balance. Development of hardware and methods for ion control will be useful for all users of recirculating hydroponic culture.

Expert Systems.

In order to incorporate all of the information about systems, plant populations, response surfaces, and mass and energy management, expert systems will be developed. Such systems could monitor environment, systems operations, and plant responses. From these inputs it could make decisions for control and issue advisories to crew for cultural needs. Expert systems could also provide a means for troubleshooting problems and issuing instructions for repairs. Such systems will be essential in plant factories and extremely useful in greenhouses, particularly those operated without resident experts.

Methods for Waste Recycling.

CELSS requires recycling of severely limited resources. While not so severely restricted, CEA production systems will face increasing pressures to conserve and recycle resources and to reduce solid wastes. Holland already restricts the disposal of growth media, inedible crop material, and nutrient salts. In other countries disposal is expensive. Methods for efficiently regenerating such wastes will be very important to the future of Controlled Environment Agriculture.

SUMMARY

Regenerative Life Support and Controlled Environment Agriculture require the same technologies for crop production. They differ only in their objectives and the constraints imposed on them by their applications. CEA is profit motivated - yield, quality, price, and costs of products are its primary concerns. Its constraints are light, climate, markets, cost-to-grow, and investment. Regenerative Life Support is concentrated on the safety and well-being of a human crew in an extraterrestrial environment. Regeneration of air and water are comparable in importance with food production. Reliability and controllability are critical. Constraints are light, environment, volume, mass, and power. The technologies developed for use in a CELSS will be immediately useful to Controlled Environment Agriculture. In fact, new products for plant production will likely result. However, as explained above, it is not new technology which drives the development of CEA. It is consumer demand in the marketplace for consistently high-quality plant products and it is the pressures of population growth and cultural change in the world. The technologies developed by NASA for CELSS will help to make it possible to market fresh plant products at reasonable prices to consumers and profit to growers, to grow fresh produce in places where nothing currently grows, and to minimize the impact of growing on the land, on resources, and on the environment.

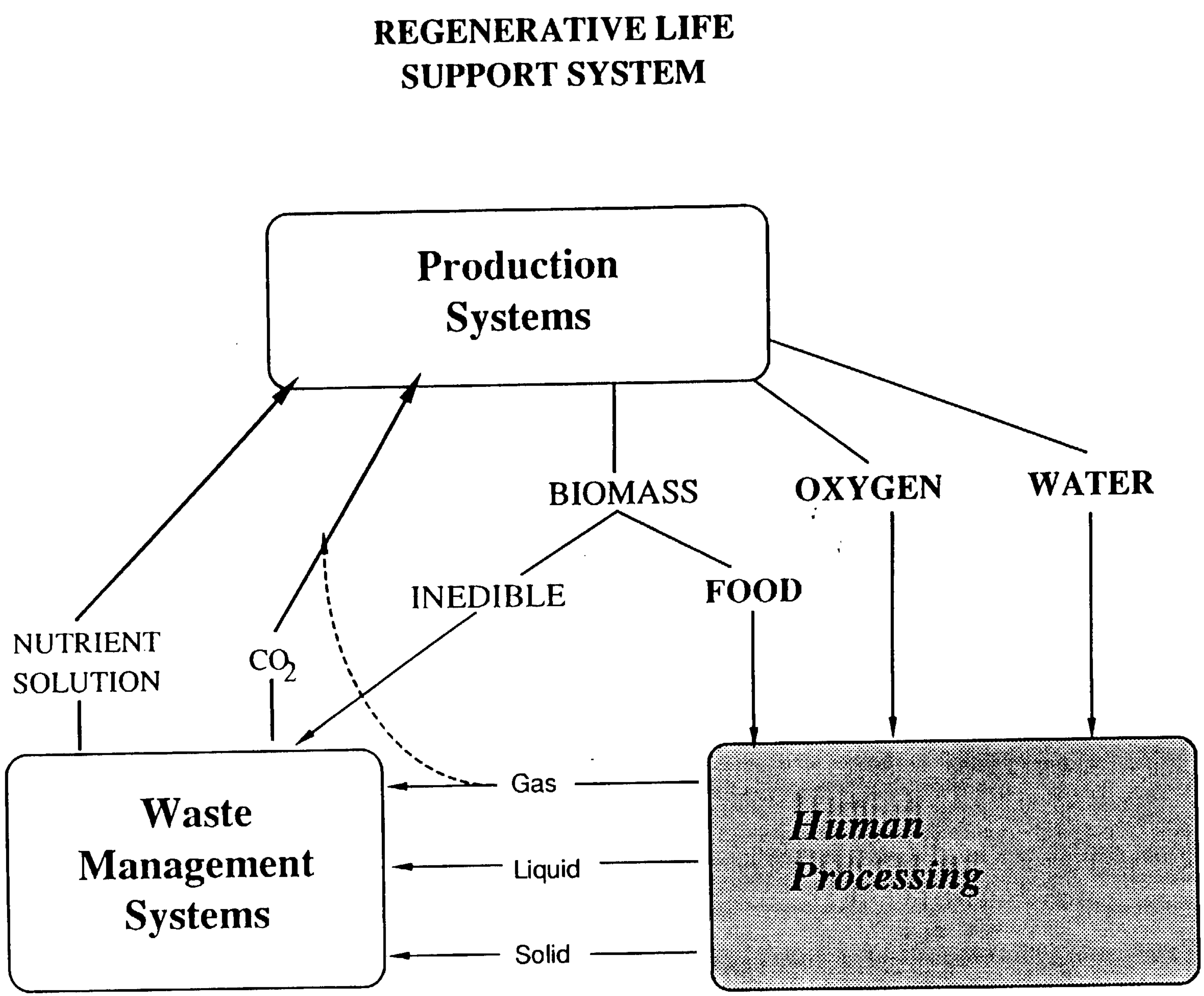


Figure 1. Diagram of the Regenerative Life Support System concept

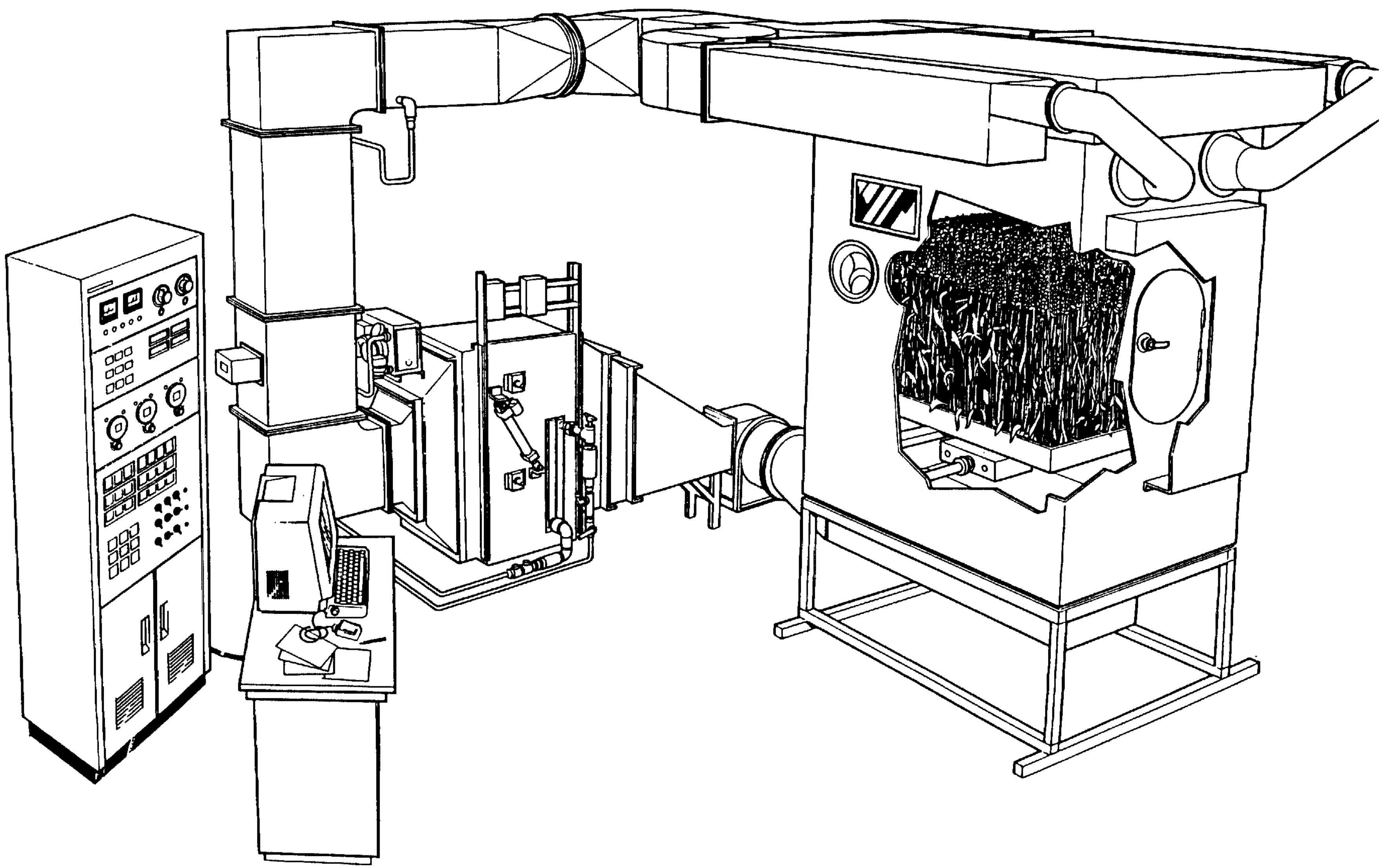


Figure 2. The Crop Growth Research Chamber is a unique tool for studying the growth of crops in controlled environments and determining their environmental response surfaces.

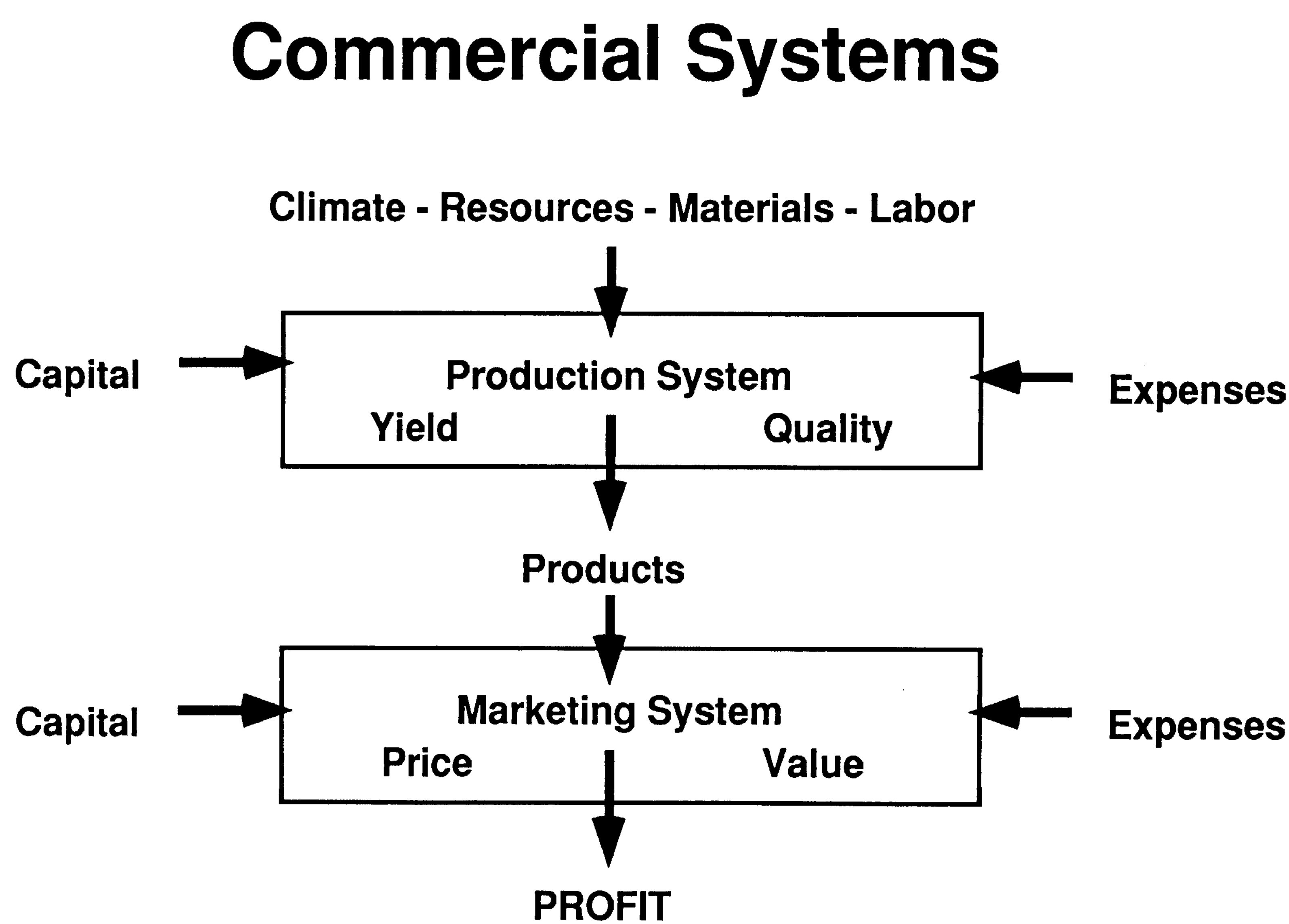


Figure 3. Diagram of commercial system. Controlled Environment Agriculture attempts to minimize effect of climate and the input of resources while maximizing yield and quality of the products grown.

	LETTUCE	TOMATOES	CUCUMBERS
FIELD CROPPING	0.5 - 3.0	0.5 - 3.0	1 - 5
GREENHOUSE	4 - 10	4 - 11	5 - 19
ABU DHABI (CEA Greenhouse)			
ACHIEVED	10.2	8.1	19.0
ESTIMATED POTENTIAL	13.8	9.0	20.7
G.E. GENIPONICS (Plant Factory)			
ACHIEVED	40.0	21.1	34.1
ESTIMATED POTENTIAL	65.0	48.0	50.0
PHYTOFARMS (Plant Factory)			
ACHIEVED	55.0		
ESTIMATED POTENTIAL	75.0		

Table 1. Yields (pounds per square foot per year) of three vegetable crops grown in various cropping systems.

N92-22481

TWO NEW ADVANCED FORMS OF SPECTROMETRY FOR SPACE AND COMMERCIAL APPLICATIONS

Kenneth J. Schlager
Biotronics Technologies, Inc.
Waukesha, WI 53186

ABSTRACT

Reagentless ultraviolet absorption spectrometry (UVAS) and Liquid Atomic Emission Spectrometry (LAES) represent new forms of spectrometry with extensive potential in both space and commercial applications. Originally developed under NASA Kennedy Space Center sponsorship for monitoring nutrient solutions for the Controlled Ecological Life Support System (CELSS), both UVAS and LAES have extensive analytical capabilities for both organic and inorganic chemical compounds. Both forms of instrumentation involve the use of remote fiber optic probes and real-time measurements for on-line process monitoring. Commercial applications exist primarily in environmental analysis and for process control in the chemical, pulp and paper, food processing, metal plating and water/wastewater treatment industries.

REAGENTLESS ULTRAVIOLET ABSORPTION SPECTROMETRY

Traditional ultraviolet absorption spectrometry requires the use of chemical reagents that produce chemical reactions that result in strongly absorbing solutions with light absorbances proportional to the target analyte concentration. These reactions often extend their effects into the visible range producing color changes in the sample solutions. The use of such reagents allows for <u>specific</u> measurements because the regent is usually specific to the analyte of interest. Reagent-based ultraviolet-visible absorption spectrometry while quite suitable for off-line laboratory use is not well adapted to on-line, real-time monitoring. The need for a continuous supply of reagents complicates the instrument design and usually results in unreliable measurements.

Many organic compounds absorb light energy in the ultraviolet region and produce characteristic "fingerprint" spectra unique to each compound. Inorganic ions, particularly the transition metals, join with other chemicals called ligands to form complexes in solution known as coordination compounds. These coordination compounds also are characterized by significant ultraviolet spectra. These ultraviolet spectra, which occur naturally without the use of chemical reagents, provide the basis for chemical analysis in combination with mathematical techniques of multicomponent chemical analysis. The specificity provided by the reagent in traditional ultraviolet spectrometry is replaced by the specificity possible with mathematical/statistical analysis.

The absorption spectrum for a particular multiple component chemical solution will be a function of the individual absorbing components in the solution and the spectral interaction of these compounds across a selected wavelength range. Several of the chemical components in nutrient hydroponic solutions (including nitrates and transition metals) are known to absorb light in the ultraviolet-visible wavelength range.

Single Component Chemical Analysis

The absorption of light in the ultraviolet-visible range is a result of shifts in the electronic energy of an atom or molecule caused by excitation from the appropriate wavelengths of light. Absorption spectrometry is based on Beer's Law which states that absorption through a liquid medium is a function of the absorptivity of the medium, the path length through the medium and the concentration of the absorbing components in the medium. These relationships can be restated to solve for concentration of an absorbing component if the absorption of a known path length through the liquid can be measured. Figure 1 illustrates the basic relationships defined in Beer's Law.

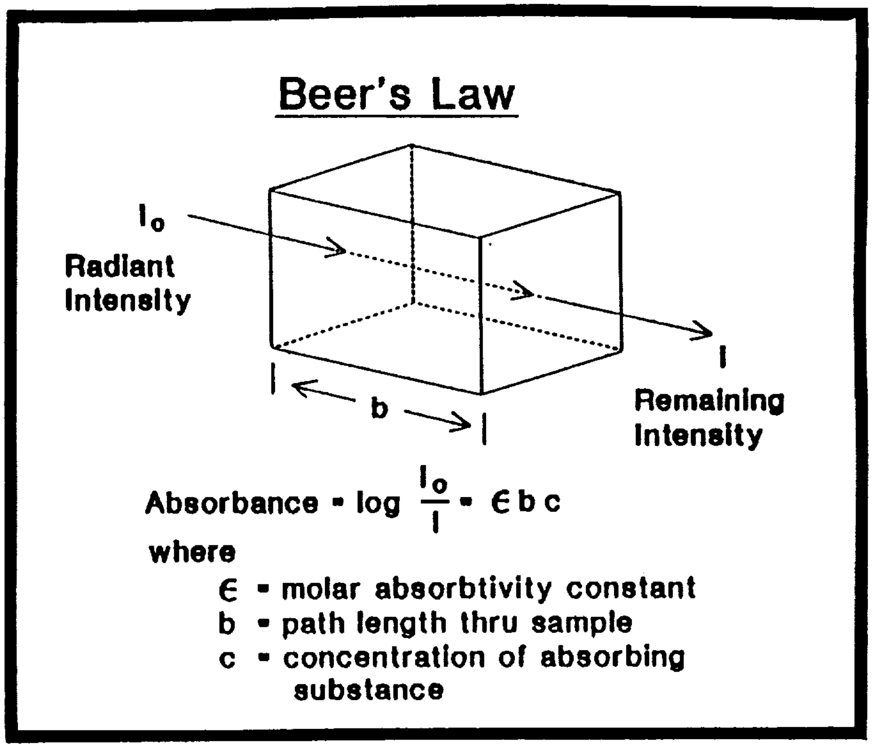


Figure 1. Beer's Law

Absorption at a specific wavelength can be thought of as the ratio of the incident intensity value of light at a specific wavelength before passing through the liquid to the value of the remaining light intensity at that same wavelength after passing through the liquid. Measurement of absorption at several wavelengths over a range will result in an absorption spectrum or "signature" for the liquid over the wavelength range with smaller intervals between wavelengths resulting in better resolution of the absorption signature.

Multicomponent Chemical Analysis

In the case of a solution with only one absorbing component, concentration can be related directly to the absorption measurements, often using a single (peak) wavelength. In more complex solutions, however, absorption at any one wavelength may be the result of the combined effects of several absorbing components. Each absorbing component has its own unique signature that remains constant in relative shape but changes in intensity as concentration changes. If this signature and the signature of the other absorbing components can be characterized, their influence on the composite absorption signature for the solution can be calculated using pattern recognition algorithms. These algorithms, frequently referred to as "chemometric models," permit absorption spectrometry to be used for multicomponent analysis.

Because specific absorbing components may contribute significantly to the absorption measurement at one specific wavelength but little or nothing to the absorption at another wavelength, multicomponent chemical analysis using absorption spectrometry will usually require measurements of absorption at several different wavelengths. If the analysis is to be performed "on-line" such as on a flowing sample, special instrumentation is needed that is capable of capturing information for a range of wavelengths and for immediate interpretation using chemometric models.

Instrumentation for Ultraviolet Absorption Spectrometry

A spectrometric instrument for ultraviolet-visible absorption spectrometry (UVAS) is shown in Figure 2. This same instrument also operates in the atomic emission mode, a form of spectrometry that is discussed in the latter part of this paper. In the absorption mode, the spectrometer transmits broad band light from a xenon flash lamp through a fiber optic cable to an optrode in a process pipe or tank. Light passes through the process fluid with selective light absorption that is dependent on the chemical composition of this same process fluid. Light surviving passage through the optrode is then returned via fiber cable to a spectrograph that disperses the light across a 200 - 800 nm spectrum. A silicon photodiode array of 1,024 elements detects light at each of 1,024 wavelengths and converts it to electrical form for subsequent processing. The 1,024 signals are sequentially converted to digital form and stored in computer memory. Spectral data is then processed through a series of signal processing and pattern recognition algorithms to estimate the chemical composition of the process. All of these measurement and processing operations are carried out in real time for on-line chemical monitoring of the process fluid.

On-line, real-time ultraviolet-visible spectrometric process analysis provides a strong capability for product improvement and cost reduction in many areas of the process industries. Some of these applications will now be discussed.

On-Line Ultraviolet-Visible Absorption Spectrometry Applications

On-line ultraviolet-visible absorption spectrometry may be applied in environmental analysis and a wide variety of process industry markets including the following:

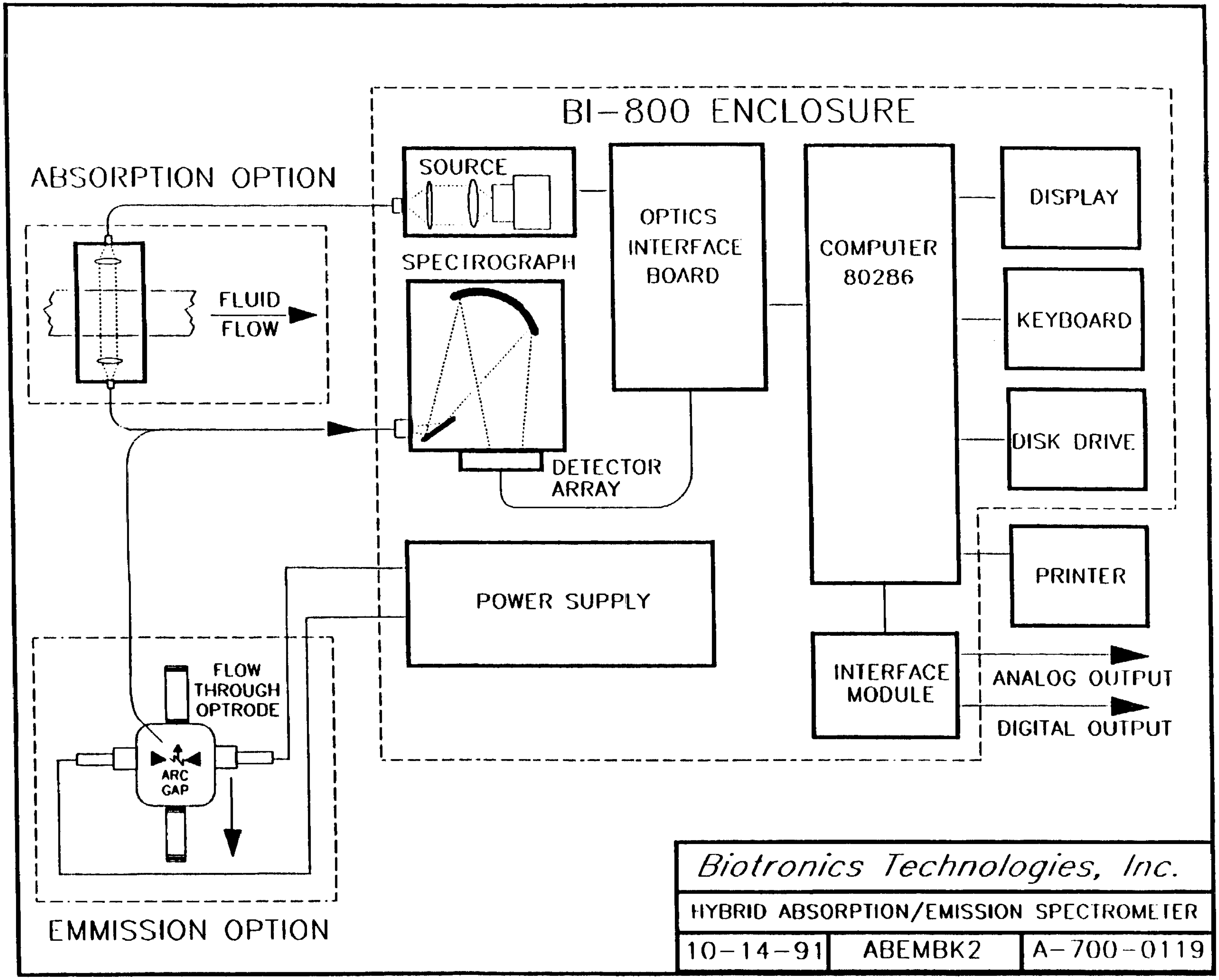
1. The chemical industry
2. The food processing industry
3. The petroleum and petrochemical industries
4. The pulp and paper industry
5. The metal plating industry
6. The electronics industry
7. The water and wastewater treatment industry

Specific applications for UVAS can be identified in each of these industries, but emphasis here will be on one specific market within the water and wastewater treatment industry: the industrial water treatment market. This market is important not only because of its specific nature but because it is a market application that is currently being field tested with Biotronics' UVAS analyzer and with production scheduled for 1992. Three primary segments exist in the industrial water treatment market:

1. In-process
2. Boiler feed
3. Cooling

UVAS at Biotronics has found its first application in the cooling water market. Chemicals are used in this market to prevent corrosion and scale (from calcium and magnesium salts) and the growth of bacteria and algae. Corrosion, scale deposition and microbiological growth can significantly reduce operating efficiency and increase plant maintenance costs. Phosphonates, azoles and molybdates are typical chemicals used in cooling water treatment. The Biotronics BI-800 UVAS instrument provides for on-line analysis of these treatment chemicals for cost economy and optimal control of corrosion and scale.

The cooling water treatment industry represents a market in excess of $2 billion annually with an estimated 1,500,000 customer sites worldwide. Biotronics' Fortune 100 chemical industry corporate partner estimates that it currently supplies 47,000 of the estimated 280,000 major sites that will have use for a water analysis instrument. The corporate partner expects to increase its market share as a result of the proprietary position offered by the new analytical instrumentation. Even without such an increase in market share, however, the cooling water instrument market represents a market potential of $235 million. Other water treatment markets such as boiler feedwater and wastewater represent future markets of equivalent size. With pending production orders scheduled for 1992, ultraviolet spectrometry has moved from a Phase I NASA SBIR in 1990 to quantity production in less than two years time.



BI-800 ENCLOSURE
ABSORPTION OPTION
SOURCE
SPECTROGRAPH
OPTICS INTERFACE BOARD
COMPUTER 80286
DISPLAY
KEYBOARD
DISK DRIVE
FLUID FLOW
DETECTOR ARRAY
POWER SUPPLY
PRINTER
INTERFACE MODULE
ANALOG OUTPUT
DIGITAL OUTPUT
FLOW THROUGH OPTRODE
ARC GAP
EMMISSION OPTION
Biotronics Technologies, Inc.
HYBRID ABSORPTION/EMISSION SPECTROMETER
10-14-91
ABEMBK2
A-700-0119

Figure 2.

LIQUID ATOMIC EMISSION SPECTROMETRY

Fourteen different primary and trace nutrients comprised the original set of CELSS requirements. With the addition of pH and dissolved oxygen, sixteen analytes were candidates for on-line measurement. These analytes are listed in Table 1. Only five of these analytes were strong candidates for effective measurement using ultraviolet-visible absorption:

1. Nitrate (NO_3)
2. Iron (Fe)
3. Manganese (Mn)
4. Copper (Cu)
5. Zinc (Zn)

Two other analytes, phosphate (H_2PO_4) and sulfate (SO_4), were weak absorbers and marginal prospects for UVAS spectrometry.

ALTERNATIVE TECHNIQUES FOR UV-VISIBLE ABSORPTION ANALYSIS

ORIGINAL STUDY OBJECTIVES			EXPANDED STUDY OBJECTIVES		
	ABSORBANCE			ABSORBANCE	
ANALYTE	PRIMARY	SECOND'Y	ANALYTE	PRIMARY	SECOND'Y
Fe	YES	YES	Na	NO	YES
K	NO	YES	Cl	NO	YES
Mg	NO	YES	Mn	YES	YES
Ca	NO	YES	Cu	YES	YES
pH	NO	YES	Zn	YES	YES
NO_3	YES	NO	MoO_2	YES	YES
H_2PO_4	YES	YES	BO_3	YES	YES
SO_4	YES	YES	O_2	NO	NO

Table 1.

An alternate approach to absorption measurements, designated as secondary absorption spectrometry, makes use of immobilized reagents on fiber optrodes. These reagents react reversibly with the analyte of interest to produce a change in absorption. This approach was demonstrated for magnesium using 8-hydroxyquinoline as the immobilized reagent and for pH with a number of reagents. Secondary absorption methods, however, raise major questions regarding the long term permanency and stability of such reagents in space or industrial environments. These questions provided a strong incentive to search for a new method of spectrometric analysis that could measure the remaining analytes of interest for NASA CELSS. This new method was discovered in a new form of emission spectrometry, Liquid Atomic Emission Spectrometry (LAES).

Atomic spectrometry is a long established technology for the analysis of gases and solids. All forms of atomic spectrometry require the application of energy to break the molecular bonds and reduce the substance to atomic form prior to analysis. Flame photometry is a traditional form of this technology still used for less demanding applications. Other forms of atomic spectrometry make use of electric arcs and sparks, graphite furnaces and radio frequency plasmas. Either light emission or light absorption techniques are used to quantify the concentrations of target analytes.

All of the above spectrometric technologies convert the material under analysis to a gaseous state or plasma prior to analysis. This state allows individual atomic elements to produce emission or absorption lines at specific wavelengths that are predictable for each element for a specific form of excitation and detection. This technique, however, requires that a sample of the substance being investigated be extracted from the source.

For on-line analysis of liquids, an analytical technique that will permit simultaneous analysis of many elements directly in the flow stream is required. Liquid Atomic Emission Spectrometry (LAES) is a new form of atomic spectrometry in which an arc or spark discharge is used to generate atomic light emission directly in the liquid medium. The light emission not only occurs directly in the liquid, but is also detected directly in the liquid using special apparatus that will permit in-situ simultaneous analysis for a spectrum of multiple elements.

Instrumentation

The instrumentation required for Liquid Atomic Emission Spectrometry is similar to the instrumentation used for UVAS. In fact, one of the principal advantages of LAES is that the same 1,024 element photodiode array detector can be used for both LAES and UVAS analysis.

The LAES instrument, originally used for experimental analysis of NASA nutrient solutions, used a beaker to hold a sample of the liquid. A flow through cell, illustrated in Figure 2, has now been developed for on-line analysis. An arc was created in the beaker using electrodes immersed in the nutrient solution and separated by an electrode gap. A fiber optic cable linked the analyzer and the beaker with the termination of the cable located in close proximity to the arc gap. This arrangement provided the energy necessary to accomplish molecular breakdown of the elements in the solution, as well as the energy necessary to stimulate atomic light emission from these elements.

The spectrometer used for the original LAES experiments was identical to that used for the ultraviolet-visible absorption experiments discussed above. The fiber optic cable was connected to a spectrograph containing a diffraction grating that separates light returned to the analyzer through the cable into discrete wavelengths from 200 - 800 nm and projects these wavelengths onto a 1,024 element linear photodiode detector array. Each element in the array is matched with a dedicated capacitor which stores spectral values that can be rapidly scanned to accumulate information and convert the information into digital values. This information can then be processed by pattern recognition software implemented in an 80286 computer built into the instrument. A block diagram of an instrument designed for both absorption and emission spectrometry is shown in Figure 2.

A series of experiments were conducted near the end of the Phase I program to explore the capability of the LAES concept and its potential as an on-line monitoring technology. Solutions were prepared using distilled water and a single compound that contained an analyte of interest. Several individual test solution were made containing known concentrations of each analyte. Each individual solution was subjected to an arc-induced excitation, and the resulting light emission was captured and recorded using the instrumentation described above. Because customized pattern recognition software was not available for this new technology, standard conventional atomic emission spectra tables were used to help identify the emission peaks observed during the experiments.

Following experimentation with the individual analyte solutions, a multicomponent nutrient solution with known components was tested, and the results were recorded.

Test Results

Individual Analyte Solutions. All of the solutions tested exhibited prominent peaks at or near the wavelengths predicted for the element under investigation when liquid atomic emission results were compared to the known emission lines from conventional atomic emission techniques. An example can be seen with sodium, shown in Figure 3, which would conventionally be expected to show peak emissions at 589 nm. The LAES experiments showed strong peaks at photodiode element number 644, which translates into a wavelength of approximately 589 nm. Furthermore, results show that the relative intensity of the peak is consistent with the relative concentration of the sodium contained in each of the test solutions. Similar results were obtained for magnesium, calcium, potassium, copper, and zinc.

Nutrient Solutions. Additional experiments were performed with a multicomponent nutrient solution in order to illustrate the ability of the LAES technique to simultaneously capture multiple element emission peaks. Although some peaks were not identified, the emission information detected was able to be matched with expected peaks for calcium, copper, zinc, hydrogen, sulfur, magnesium, molybdenum, manganese, potassium and oxygen, as shown in Figure 4. No attempt was made to quantify these elements due to the lack of sufficient information to establish adequate calibration models.

Liquid Atomic Emission Applications

Because Liquid Atomic Emission Spectrometry is still in the development stage, it is not yet possible to speak of demonstrated applications ready for commercial production. Nevertheless, two major applications should be in the field testing stage by the middle of 1992: environmental monitoring of toxic metals and process monitoring and control in the metal plating and electronics industries.

Toxic metals such as lead, chromium, nickel and cadmium are a recognized environmental health threat. There is a pressing need for on-line monitoring of these metal analytes in drinking water, wastewater and in on-site soil surveys. Biotronics is currently working with the U.S. Army Corps of Engineers in developing an on-site system for direct detection and quantification of toxic metals in soils. As an augmentation of the Army Corps of Engineers SCAPS (Site Characterization and Analysis Penetrometer System), this system would provide on-site three-dimensional mapping of toxic metals in survey areas at depths of up to 150 feet. This same LAES should find an extensive market in municipal and industrial wastewater treatment.

Metal plating is a process used in a variety of industries from jewelry and automobiles to electronics (integrated circuits and printed circuit boards) and aerospace. Both quality and cost control of metal plating processes could significantly benefit from on-line monitoring and control metal concentrations in plating baths. This multi-billion dollar market is an ideal one for LAES because the new spectrometric technology is capable of direct plating bath measurement of metal concentrations. Biotronics has previous experience in this industry and is already working with a major supplier of plating solutions to develop this market application.

Phase I SBIR to Market in Two Years

Biotronics' on-line BI-800 ultraviolet/visible spectrometer represents a NASA success story in technology transfer. The compressed time period of the transition from research concept to volume production is evidenced from the following timetable of events:

Phase I SBIR Completed - July 3, 1990
Corporate Partner Development Agreement Executed - September 12, 1990
First Prototypes Delivered to Corporate Partner - June, 1991
First Production Spectrometers Delivered - March, 1992

It is important to emphasize that the above development scenario includes only the UVAS portion of the NASA SBIR program. A Phase II contract to develop the more experimental Liquid Atomic Emission Spectrometer (LAES) was concluded in May, 1991. Market implementation of this technology is approximately two years away. Nevertheless, the fact remains that a NASA sponsored Phase I SBIR has resulted in a commercial product in less than two years from the completion of the Phase I feasibility study.

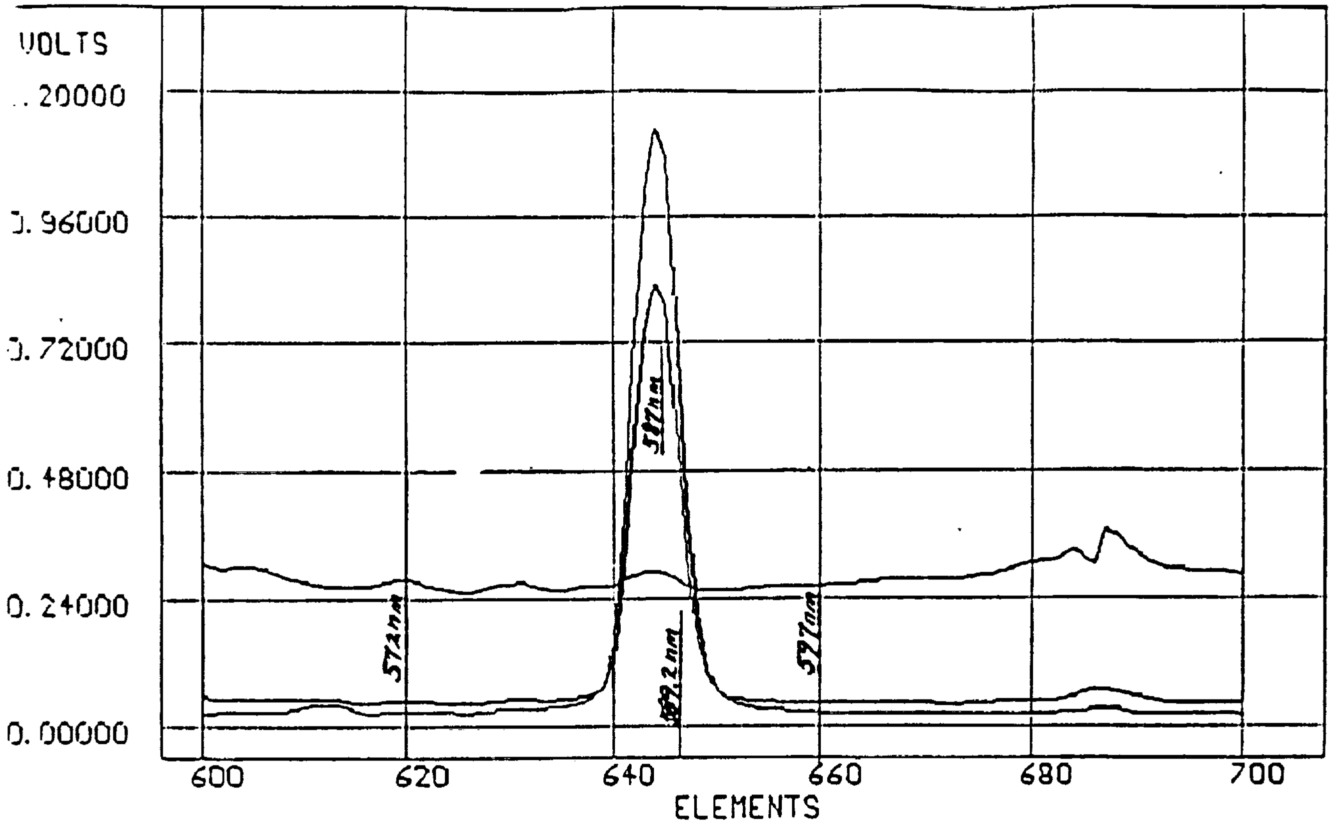


Figure 3. Atomic Emission Spectra for Sodium

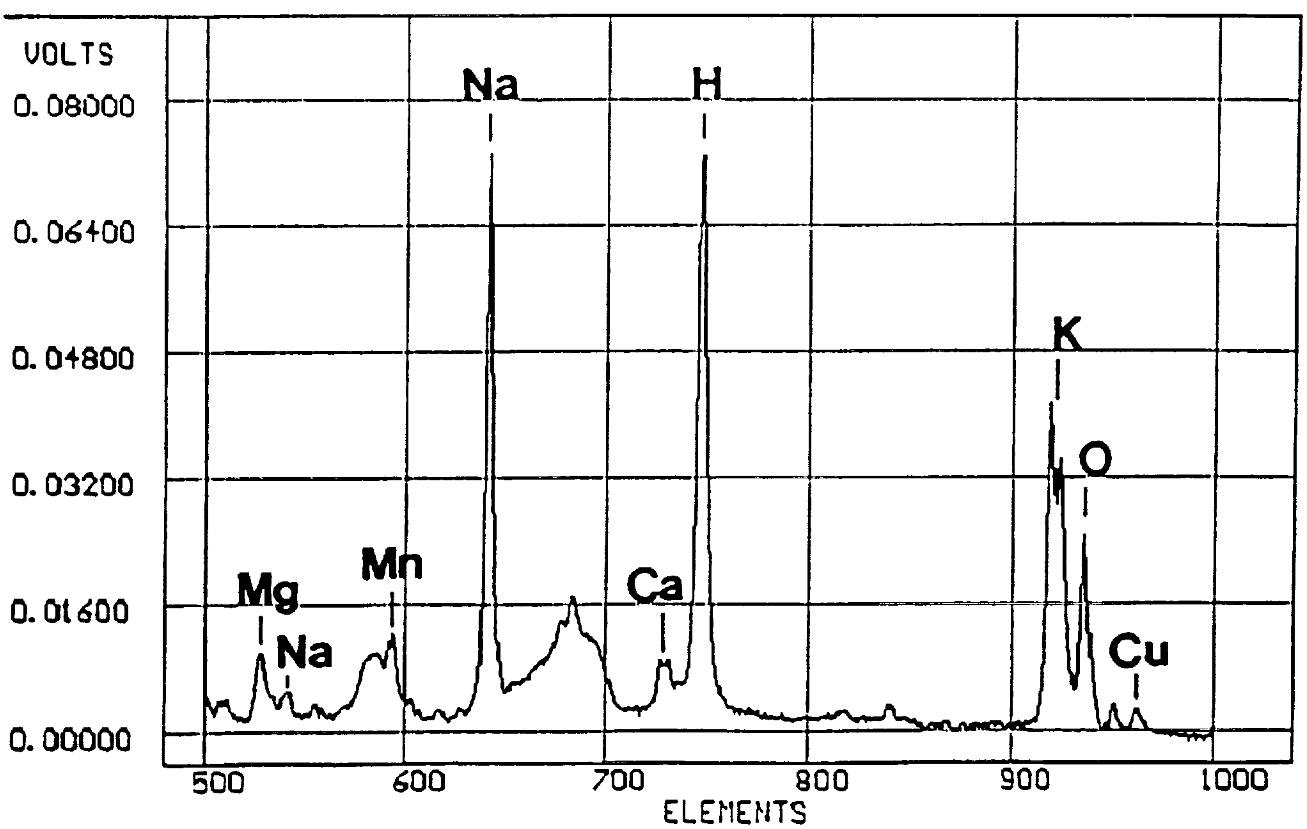


Figure 4. Atomic Emission Spectrum for a Nutrient Solution

N92-22482

A COATED-WIRE ION-SELECTIVE ELECTRODE FOR IONIC CALCIUM MEASUREMENTS

John W. Hines, MSEE, and Sara Arnaud, MD
NASA-Ames Research Center
Moffett Field, California

Marc Madou, PhD, Jose Joseph, PhD, and Arvind Jina, PhD
Teknekron Sensor Development Corporation
Menlo Park, California

ABSTRACT

A coated-wire ion-selective electrode for measuring ionic calcium has been developed, in collaboration with Teknekron Sensor Development Corporation (TSDC), Menlo Park, CA. This coated wire electrode sensor makes use of advanced, ion-responsive polyvinyl chloride (PVC) membrane technology, whereby the electroactive agent is incorporated into a polymeric film. The technology greatly simplifies conventional ion-selective electrode measurement technology, and is envisioned to be used for real-time measurement of physiological and environmental ionic constituents, initially calcium. A primary target biomedical application is the real-time measurement of urinary and blood calcium changes during extended exposure to microgravity, during prolonged hospital or fracture immobilization, and for osteoporosis research. Potential Advanced Life Support applications include monitoring of calcium and other ions, heavy metals, and related parameters in closed-loop water processing and management systems. This technology provides a much simplified ionic calcium measurement capability, suitable for both automated *in-vitro*, *in-vivo*, and *in-situ* measurement applications, which should be of great interest to the medical, scientific, chemical, and space life sciences communities.

INTRODUCTION

Measurement of chemical and organic constituents is a critical requirement for many hospitalized patients. Parameters such as ionic calcium (Ca^{2+}), potassium (K^+), and blood gas constituents (pH, PO_2, PCO_2) are of great interest for determining the clinical and physiological implications of illness and various therapies as well as long-term spaceflight missions. As part of NASA's planned Life Sciences and Advanced Life Support Space Technology Programs, we are developing minimally invasive technologies which can be used to measure important ionic constituents on Earth and during prolonged spaceflight. This technology can also be applied to monitor and control such necessities as water and air quality, and to effect closed-loop control on reclamation and management of those processes.The calcium CWE represents the initial demonstration of this technology.

Ion-selective electrodes are becoming increasingly important in measuring ion concentrations in solutions. The continuing study of and application of liquid and polymer membrane ion-selective electrodes (ISE) remains a strong and viable area of interest in several laboratories. The theory behind the operation of ion-exchange membranes has been investigated and developed on the basis of the concept of zero-current potential. Ion-selective electrodes have made great inroads into the testing of electrolytes and have replaced flame photometers for this use in many laboratories. Ion selective electrodes have found application in the control of industrial processes, water supplies and waste water. In addition, they are used extensively in biomedical applications. Of the liquid membrane-based electrodes incorporating ion-exchange materials, calcium-selective electrodes in particular continue to draw interest. Much of the work has been directed to the way in which electrodes are constructed. The desire to miniaturize, simplify and to produce cheaper ion-selective electrodes has engrossed a diverse group of research workers.

Presently no reliable sensors exist for undertaking real-time *in vitro*, *in vivo* , and *in-situ* measurements during space flight missions. The development of such microsensors will permit real time monitoring of astronaut health and especially, gradual degenerative changes such as bone mass loss, which occurs at or near zero gravity. The development of a suitable calcium microsensor for medical and physiological measurements will not only have profound applications for NASA but also for the study of osteoporosis - a disorder which affects many postmenopausal women.

Advanced Life Support requirements for such planned mission scenarios as Lunar and Martian bases have placed stringent demands on sensors, and measurement and control technology for closed-loop physical/chemical water and air quality systems. The list of required and potentially required parameters include ionic constituents, heavy metals, detergents, microbes, organics and trace contaminants, and microbial constituents, such as bacteria and viruses. The Coated-Wire Calcium Electrode is but one possible method for measuring such constituents, but represents and important approach due to the simplicity of the configuration.

Conventional ion-selective electrodes

The conventional cell arrangement for potentiometric measurements using an ion-selective electrode or membrane electrode is shown in figure 1, below.

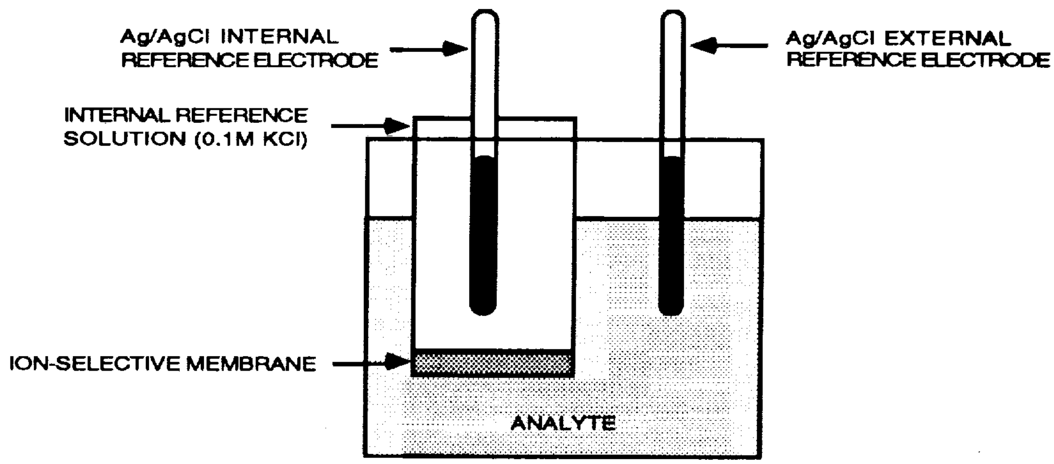


FIGURE 1: EXAMPLE OF A CONVENTIONAL ION-SELECTIVE ELECTRODE

In this configuration, the internal aqueous reference solution, which contains the ions necessary to maintain a constant potential at both the reference element and the inner surface of the ion -selective membrane is held at a constant composition. The potential difference across the cell is dependent on the composition of the sample solution, and the specific ion that will be detected is determined by the nature of the membrane.

Coated wire ion-selective electrodes

The coated wire electrode (CWE) uses components of conventional ion selective electrodes except that no internal aqueous filling solution is used (figure 2). Instead a conductor is directly coated with an ion-responsive membrane, usually polyvinyl chloride (PVC) based. Thus in a coated wire electrode, the electroactive agent is incorporated into a polymeric film. The conductor can be metallic or graphite-based and be of any convenient geometric shape (i.e. wire, disk, cylinder, thin film, etc.). When the ion-selective membrane comes into contact with an ionic solution an ion exchange reaction takes place resulting in a potential difference across the solution-membrane interface. In coated-wire electrodes the cell configuration is as follows:

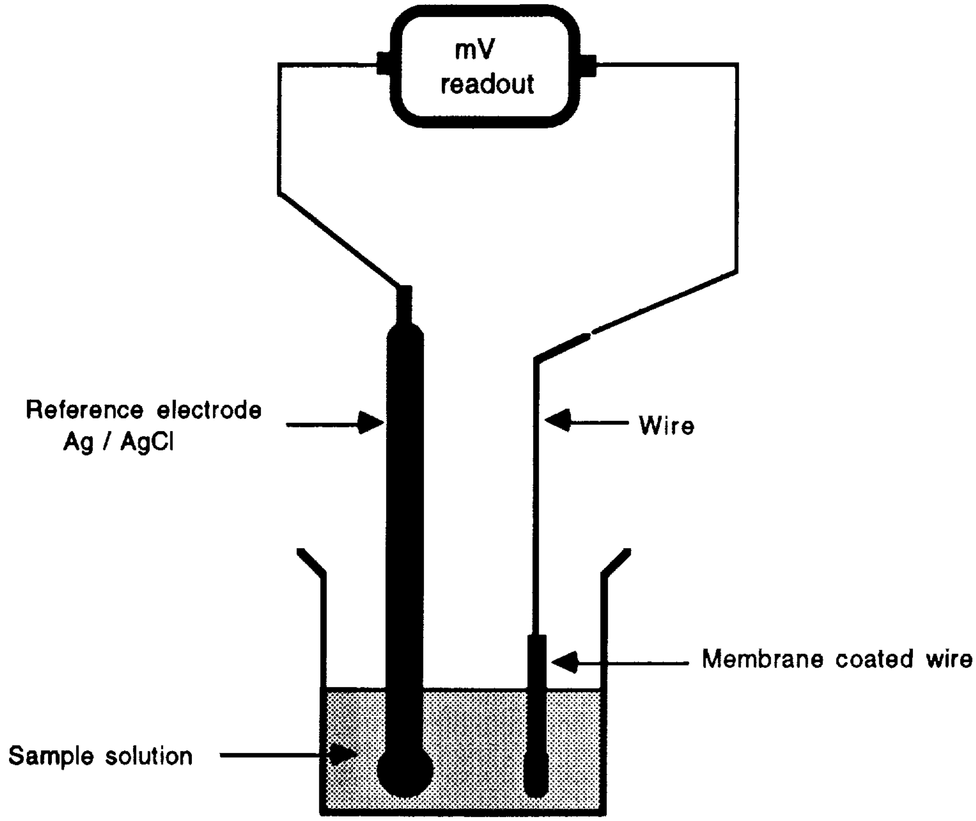


FIGURE 2: EXAMPLE OF A COATED WIRE ELECTRODE

EXPERIMENTAL METHODS

Initial Development Phase

The general procedure for making a Coated Wire Electrode is given below:

1. Preparation of the membrane solution

The membrane solution for a Ca2+ ion-selective electrode was prepared using an optimized formulation consisting of a Calcium ionophore, specific binding and reactive agents, and polyvinyl chloride.

2. Application of membrane to silver wire

A straight piece of silver wire 8 cm long, previously cleaned, was dip-coated several times by immersing it into the above membrane solution. Great care was exercised to prevent the occurrence of any pin-holes by ensuring that a relatively thick membrane was coated on to the wire as shown in figure 3 below:

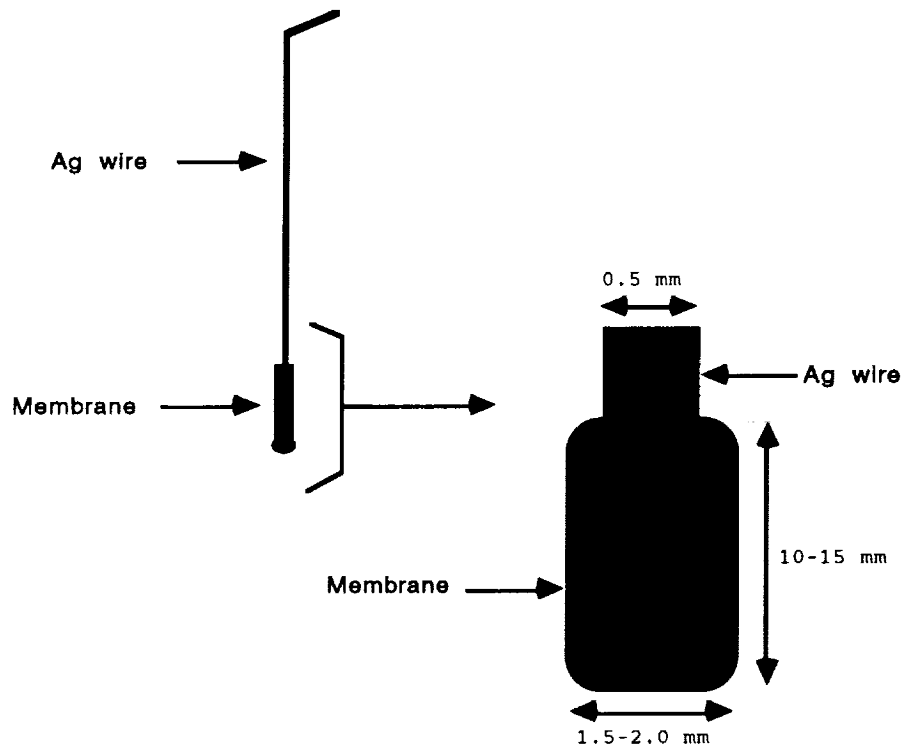


FIGURE 3: REPRESENTATION OF A CALCIUM ION SELECTIVE MICROELECTRODE

In applying a thick membrane coating, the main aim is to provide good adhesion to the metal surface and to obtain a film devoid of pin holes. It should be noted that although the 'dip-coating' procedure applies specifically to wires, it can also be adapted for other configurations such as plates and discs. If necessary the remaining exposed metal surface of the electrode can be insulated by tightly wrapping parafilm around it. After application of the membrane coat, the electrodes were allowed to dry for several hours or overnight before measurements were taken. In some instances the electrodes were stored dry for three weeks before taking measurements. Prior to taking any readings the electrodes were allowed to equilibrate in a dilute calcium chloride solution for about an hour. This precaution is usually only needed initially, since subsequent equilibration times are much shorter (a few minutes).

3. Initial testing of electrodes

All measurements and tests were conducted using the experimental setup illustrated in figure 4 below:

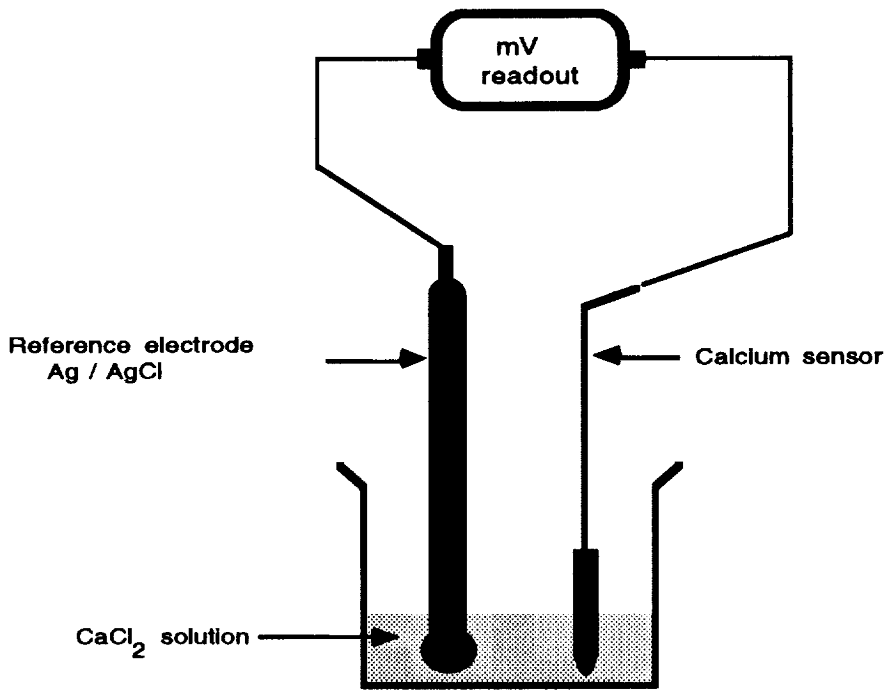


FIGURE 4: EXPERIMENTAL SETUP USED FOR TESTING CALCIUM WIRE ELECTRODES.

An external Ag/AgCl electrode was used as a reference electrode. Both the calcium ion selective electrode and the Ag /AgCl reference electrode were connected to an Orion pH / mV monitoring unit as shown above. For this initial evaluation, a set of standard $CaCl_2$ solutions ranging in concentration from 1×10^{-1} - 1×10^{-5} M and containing 0.1 M KCl were prepared, lacking any chelating agents such as EDTA or EGTA, to determine the slope and response of the microelectrodes.

EXPERIMENTAL RESULTS

Under ideal conditions, the potential difference across the solution-membrane interface is given by the Nernst equation:

$$E = \frac{RT}{ZF} \ln a_{Ca^{2+}},$$

where **Z** and **a** are the charge and activity of the ion, **R** is the gas constant, **T** is the temperature and **F** is the Faraday constant. The theoretical sensitivity as predicted by the Nernst equation is approximately 58 mV/pH unit at ambient temperature for a monovalent cation and about 29 mV/pH unit for a divalent cation.

Several of the Ca^{2+} ion selective electrodes were tested using the experimental setup shown in figure 4. Typical response curves for four such electrodes are shown in figure 5 below:

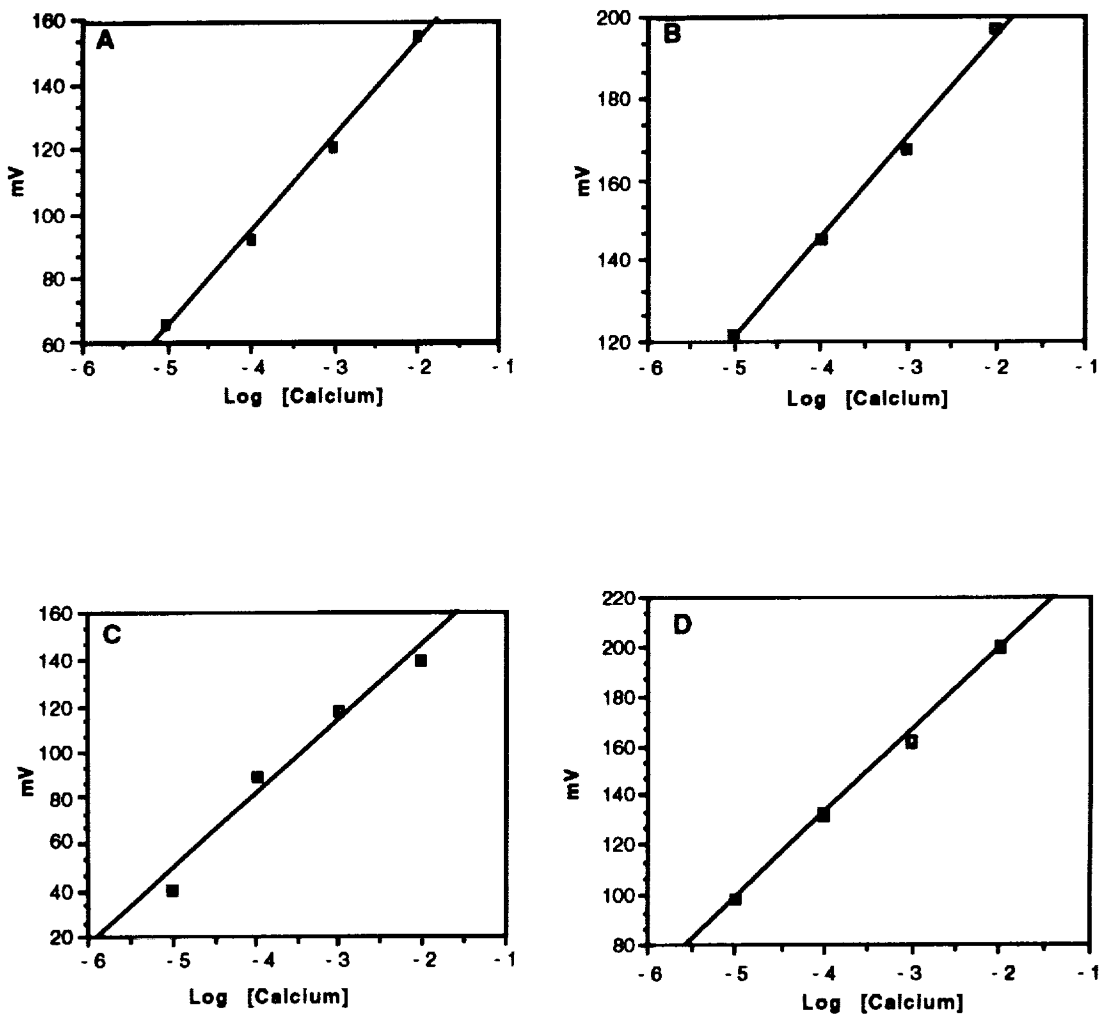


FIGURE 5: RESPONSE CURVES OF CALCIUM ION SELECTIVE MICROELECTRODES

The tests were conducted by measuring the responses progressively in increasing concentrations of $CaCl_2$ solutions. Although the initial potentials of the above electrodes in 1 x 10^{-5} M $CaCl_2$ solution varied, the average decade change between solutions is about 26 mV which compares favorably with the predicted theoretical value of a divalent cation according to the Nernst equation. When the electrodes were immersed in solution for 2 - 4 hours the electrode potentials displayed negligible drift, thus indicating that the Ca^{2+} ion selective microelectrodes have relatively stable potentials. However, it should be noted that, at this stage in the development, the construction of these electrodes is much of an art and therefore it is quite possible that microelectrodes made from the same batch could vary considerably in their properties.

Comparison of the Coated Wire Electrode with Reference Instrumentation

After having established the slope of the response of the coated wire sensors, we undertook a study to determine the compatibility of the Ca^{2+} selective membrane with the standard, commercially available Radiometer electrode assembly. An enlarged top view is presented in Figure 6.

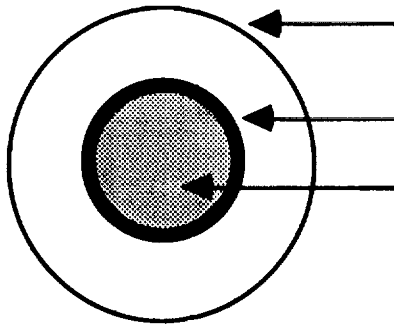


Figure 6: Top View of Coated Membrane

The original membrane on the Radiometer electrode was carefully removed and the surface cleaned with water and isopropanol. The calcium selective membrane was then immobilized onto the clean surface. A thick coating was applied to ensure a membrane devoid of pin holes. The immobilized membrane was then allowed to air-dry overnight. The electrode was reassembled as described by the manufacturer (Radiometer Copenhagen). We experienced great difficulty in removing air bubbles that were trapped in the jacket containing the electrolyte solution. The air bubbles prevent intimate contact between the membrane and the internal electrolyte solution. We believe the present design of the Radiometer electrode can be improved considerably to avoid these problems. To ensure reasonable contact between the internal electrolyte solution and the membrane, a short wick (made of cotton thread) was introduced. Two such electrodes were prepared and reassembled.

Both reassembled Radiometer electrodes containing the calcium ion selective membrane were then tested initially using standard $CaCl_2$ solutions (as shown in Figure 2) with the Radiometer Model ICA1 Ionized Calcium Analyzer. During initial , standalone testing , we observed a somewhat linear response to increasing $CaCl_2$ solutions. Due to excessive signal noise, it was not possible to obtain steady readings. However, when the electrodes were tested with the Radiometer Model ICA1 Ionized Calcium Analyzer, we obtained a very good response to various known concentrations of $CaCl_2$ solutions. The response of these electrodes compared very favorably with previous results obtained using the standard, commercial electrode system. For example, after calibration of these electrodes with the calibration solutions provided by the manufacturer the following results were obtained:

TSDC Electrode #1

		Measured	Expected
Solution A	Ca^{2+} concentration mmol/l	1.26	1.25
Solution B	Ca^{2+} concentration mmol/l	0.75	0.69 - 0.83
Solution C	Ca^{2+} concentration mmol/l	1.75	1.69 - 1.83

TSDC Electrode #2

		Measured	Expected
Solution A	Ca^{2+} concentration mmol/l	1.28	1.25
Solution B	Ca^{2+} concentration mmol/l	0.79	0.69 - 0.83
Solution C	Ca^{2+} concentration mmol/l	1.82	1.69 - 1.83

SUMMARY AND CONCLUSIONS

These results strongly indicate that the Calcium ion selective membrane is compatible with the commercial Radiometer electrodes. In addition, this membrane meets the requirements of sensitivity and selectivity to develop a Microcalcium Sensor for *in-situ* and *in vivo* measurements of ionic calcium. During the next phase of this project, we plan to evaluate the long-term stability of the ion selective membrane, and expand the technology demonstrations into more rigorous feasibility studies, using representative, physiological, clinical, and process control water stream samples. We intend to continue the development efforts to a usable prototype level, which can then be used in research, testbed, and operational spaceflight mission development activities. The Coated-Wire Electrode,when coupled with appropriate measurement instrumentation, has potential commercial application as a small, self-contained, portable instrument, capable of real or near real-time usage.

N92-22483

A 99% PURITY MOLECULAR SIEVE OXYGEN GENERATOR

G. W. Miller
Crew Technology Division
Armstrong Laboratory
Brooks AFB, San Antonio, Texas 78235-5301

ABSTRACT

Molecular Sieve Oxygen Generating Systems (MSOGS) have become the accepted method for the production of breathable oxygen on military aircraft. These systems separate oxygen from aircraft engine bleed air by application of pressure swing adsorption (PSA) technology. Oxygen is concentrated by preferential adsorption of nitrogen in a zeolite molecular sieve. However, the inability of current zeolite molecular sieves to discriminate between oxygen and argon results in an oxygen purity limitation of 93-95% (both oxygen and argon concentrate). The goal of this effort was to develop a new PSA process capable of exceeding the present oxygen purity limitations. A novel molecular sieve oxygen concentrator was developed which is capable of generating oxygen concentrations of up to 99.7% directly from air (U.S. Patent No. 4,880,443). The process is comprised of four adsorbent beds, two containing a zeolite molecular sieve and two containing a carbon molecular sieve. This new process may find use in aircraft and medical breathing systems, and industrial air separation systems. The commercial potential of the process is currently being evaluated.

INTRODUCTION

Molecular sieve oxygen generating systems are replacing liquid oxygen systems as the principal method for the production of breathable oxygen on-board military aircraft. The oxygen-rich product gas is breathed by the aircrew for the prevention of hypoxia at high altitudes. When compared to conventional liquid oxygen (LOX) systems, MSOGS systems offer many benefits, such as, reduced life cycle cost, reduced logistic support, increased aircraft versatility, and improved safety. Presently, the U.S. Air Force has several MSOGS-equipped aircraft, the F-15E "Strike Eagle," and the B1-B and B-2 strategic bombers.[1] Also, the U.S. Navy has several MSOGS-equipped aircraft, such as, the AV-8B. In the future nearly all U.S. military aircraft will be equipped with an MSOGS breathing system.

MSOGS breathing systems are comprised primarily of a molecular sieve oxygen concentrator (or generator) and associated equipment for distribution and delivery of the product gas, such as, breathing regulators. The critical component of the system is the oxygen concentrator which separates oxygen from the aircraft engine bleed air (compressed air) by pressure swing adsorption technology. Using this technology, nitrogen is preferentially adsorbed in the molecular sieve at moderate pressures, thereby, concentrating oxygen. Subsequently, the nitrogen is released to the ambient atmosphere as a waste gas and the oxygen is breathed by the aircrew. Control of the oxygen concentration is accomplished by either diluting the product gas with cabin air or by varying one of the concentrator operating parameters, such as, cycle time. The concentrator need only be supplied engine bleed air and a small amount of electrical power to produce a continuous stream of concentrated oxygen.

An oxygen generator based on the current technology is comprised of two zeolite molecular sieve adsorbent beds, several valves, a purge orifice, and an electronic timer (Figure 1). The electronic timer controls the opening and closing of the valves. Bleed air pressure is typically in the range of 137.9 to 344.8 KPa (20 to 50 psig) (referenced to the ambient atmospheric pressure). The pressure swing adsorption technique is achieved by alternating the pressurization of the two adsorbent beds. While one bed is pressurized, the opposite bed is depressurized and exhausts previously adsorbed gases to the surrounding atmosphere. In an aircraft this exhaust gas is vented overboard. In Figure 1, the cycle begins by the simultaneous opening of valves V1, V4, and V6. In this step of the cycle bed A is pressurized and nitrogen in the feed air is adsorbed in the molecular sieve. Hence, oxygen is concentrated and withdrawn through V1.

Simultaneously, bed B is vented to the surrounding atmospheric pressure through valve V6 and purged by a countercurrent flow of product gas entering through the purge orifice. In the second half-cycle valves V2, V3, and V5 are opened and valves V1, V4, and V6 are closed. The beds simply reverse roles, whereby, bed B produces the product gas and bed A is vented. A typical concentrator has a cycle time (duration of pressurization and depressurization) of 10 seconds. By alternating the opening and closing of the two sets of valves, a continuous stream of concentrated oxygen is produced.

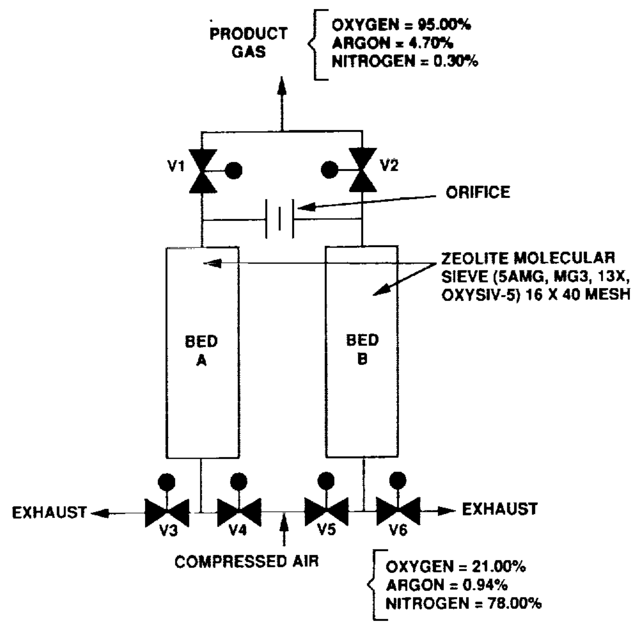


Figure 1. A Standard Two-bed Molecular Sieve Oxygen Generator.

Current molecular sieve oxygen concentrators use adsorbent beds containing exclusively zeolite molecular sieves. Several varieties (5AMG, MG3, 13X, and OXYSIV-5) are commercially available, however, most oxygen concentrator manufacturers presently use either 5AMG or OXYSIV-5 (products of UOP, Des Plaines, IL). Zeolite molecular sieves are synthetic alkali metal aluminosilicates which have as their basic building blocks SiO_4 and AlO_4 tetrahedra with exchangeable cations. The type of crystal framework and exchangeable cation will determine the dimension of the crystal pores. 5AMG molecular sieve has Type 5A zeolite crystallites with uniform pore openings of 4.2 Angstroms. MG3, 13X, and OXYSIV-5 have 13X zeolite crystallites with pore openings of 7.4 Angstroms. These materials are generally stable at high temperature but slowly deactivate in the presence of water due to the water molecule's small kinetic diameter and high polarity. Because nitrogen, oxygen, and argon have molecular kinetic diameters of 3.64, 3.46, and 3.40 Angstroms, respectively, these gases readily enter the 5A and 13X crystallites.[2] Separation of oxygen and nitrogen is possible because of a difference in equilibrium adsorption capacity (Figure 2). Molecular sieves adsorb greater quantities of nitrogen than oxygen due to the nitrogen molecule's slight polarity. Oxygen and argon concentrate because zeolite molecular sieves are unable to discriminate between these molecules. This characteristic is verified by the nearly identical oxygen and argon equilibrium adsorption isotherms in Figure 2. Both oxygen and argon are nonpolar and have nearly identical kinetic diameters. Hence, the maximum oxygen concentration from current oxygen concentrators is constrained at 95% (the remainder is mostly argon with less than 1% nitrogen). Further, an oxygen concentration of 95% is only produced under the most ideal conditions. In general, the most probable concentration limit is 93% oxygen (the remainder is mostly argon with 1-3% nitrogen).

The 95% oxygen concentration constraint is one limitation of the present oxygen concentrator technology. Conventional liquid oxygen systems have routinely supplied oxygen at a concentration of 99.5%. Although these higher concentrations may not always be required, situations do occur while

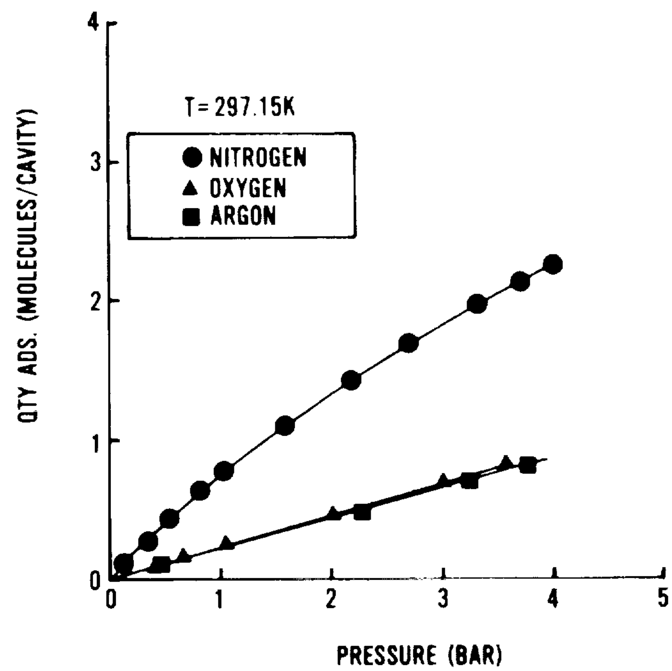


Figure 2. Equilibrium Adsorption Curves for 13X Molecular Sieve.

flying where breathing very high purity oxygen (≥99%) would be desirable. Current MSOG systems generally require a bottled backup oxygen system pressurized with 99.5% oxygen. Although MSOGS technology offers many advantages over conventional liquid oxygen systems, this technology is presently unable to produce the oxygen purity which has been routinely available from the conventional liquid oxygen systems. Hence, the goal of this work was to develop a process based on PSA technology which is capable of exceeding the 93-95% oxygen concentration constraint of current technology.

ADSORPTION BREAKTHROUGH STUDIES

In 1986 an effort was initiated with the goal of identifying an adsorbent capable of discriminating between oxygen and argon. Because of the very slight differences in adsorption characteristics between oxygen and argon molecules, the probability of finding an adsorbent with the proper characteristics was considered low. Several commercially available adsorbents (zeolite molecular sieves, such as, 3A, 4A, 5AMG, 13X, and MG3, and carbon molecular sieves) were evaluated by analysis of adsorption breakthrough curves. It was hypothesized this ability to discriminate between oxygen and argon would manifest itself as a shift in the oxygen and argon concentration wavefronts at the outlet of a single adsorbent bed. Further, it seemed logical that the oxygen wavefront must

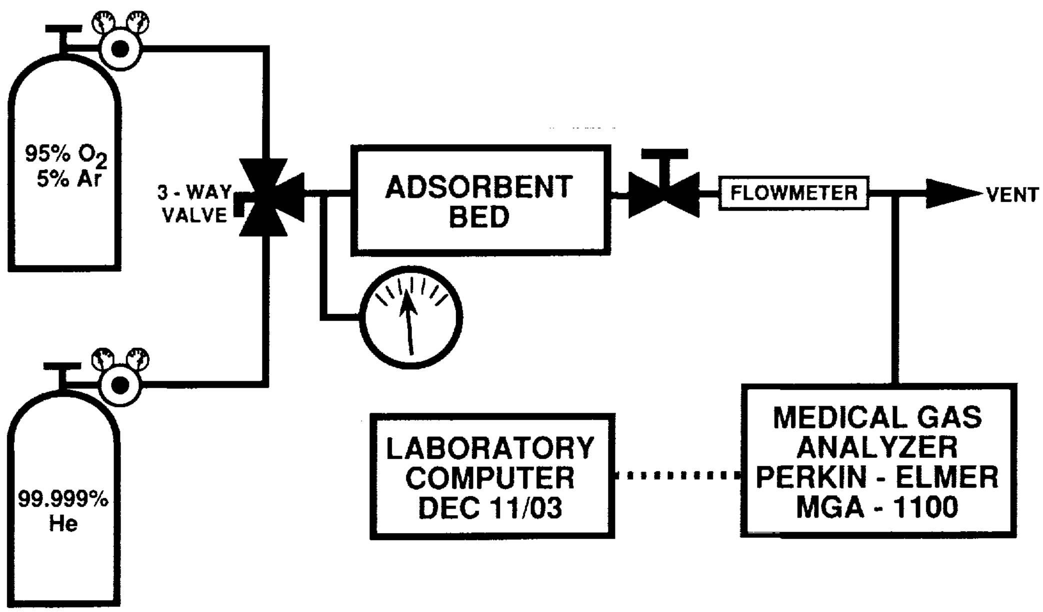


Figure 3. Adsorption Breakthrough Apparatus.

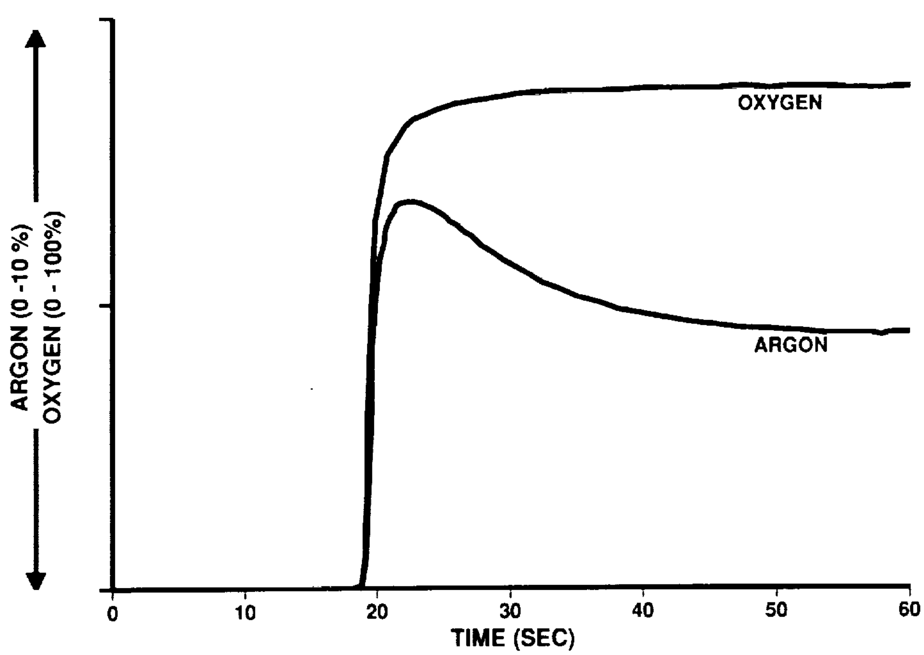


Figure 4. Adsorption Breakthrough Curves for 13X Molecular Sieve.

lead the argon wavefront, if oxygen is to be extracted as the final product from an operating oxygen concentrator.

In these experiments a single bed or column was filled with the candidate adsorbent and flushed with helium (Figure 3). Helium is used as the purge gas because it adsorbs in negligible quantities. A gas mixture with a concentration of 95% oxygen and 5% argon was then allowed to flow through the bed. This gas mixture was selected because it was assumed the product from a conventional oxygen concentrator would become the feed gas to a final oxygen purifier. The wavefronts exiting the bed were monitored by a mass spectrometer (Perkin-Elmer Medical Gas Analyzer, Model No. MGA-1100). If the oxygen and argon wavefronts overlapped upon exiting the bed, the adsorbent was considered unable to discriminate between oxygen and argon (Figure 4). However, if a noticeable shift in the wavefronts occurred, it was assumed the adsorbent could discriminate between oxygen and argon. Based on a qualitative analysis of the data, only a carbon molecular sieve caused a shift in the oxygen and argon wavefronts (Figure 5). The next step was to configure a PSA apparatus with carbon molecular sieve beds, such that, this shift in wavefronts could be effectively applied for the further purification of oxygen from a conventional oxygen concentrator.

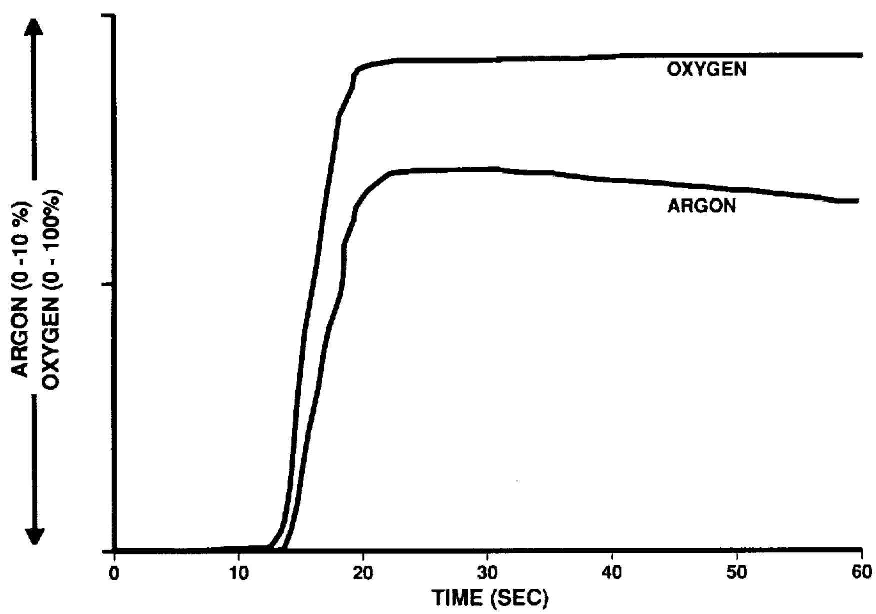


Figure 5. Adsorption Breakthrough Curves for Carbon Molecular Sieve.

SECONDARY OXYGEN PURIFIER

In 1987 a small-scale device, referred to as the "secondary oxygen purifier," demonstrated that further purification of the product gas from a standard oxygen concentrator was possible (Figure 6).[3,4] The device was comprised of two carbon molecular sieve adsorbent beds, several valves, and an electronic timer. Operation of the valves was identical to that of a standard oxygen concentrator. However, the device did not possess a purge orifice (Figure 1). The elimination of the purge orifice improved the performance of the unit and reduced the inlet gas consumption. The apparatus was fed a bottled gas with a composition of 94.73% oxygen, 5.00% argon, and 0.27% nitrogen which simulated a standard oxygen concentrator product gas. During the pressurization step argon preferentially adsorbs on the carbon molecular sieve, thereby, increasing the purity of the oxygen in

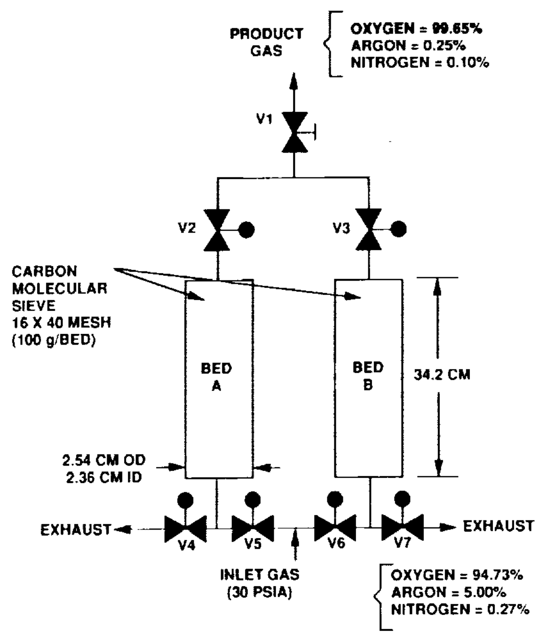


Figure 6. A Small-scale Secondary Oxygen Purifier.

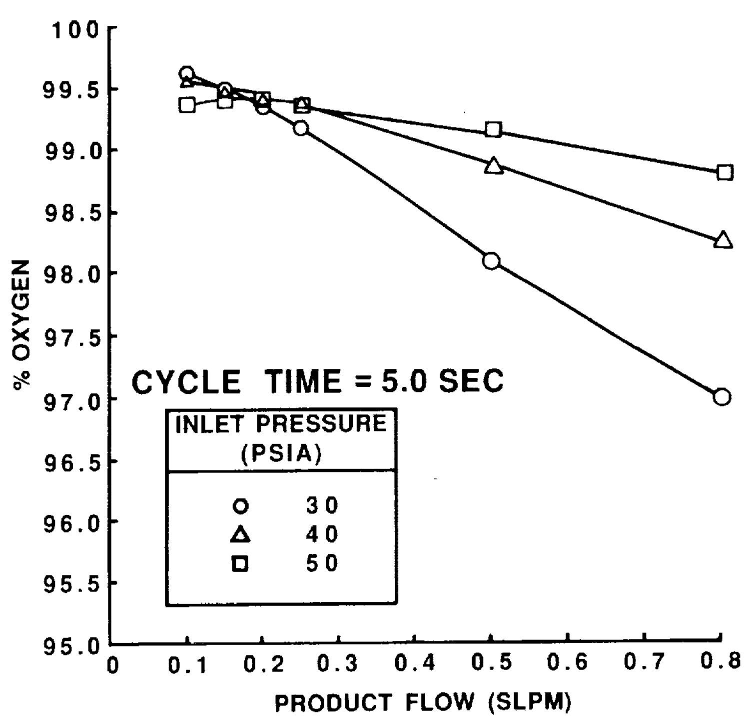


Figure 7. Performance Curves for the Small-scale Secondary Oxygen Purifier.

the gas phase. During depressurization the argon is exhausted to the surrounding atmosphere. The device produced a product gas with a concentration of 99.65% oxygen, 0.25% argon, and 0.10% nitrogen. Also, the apparatus produced nearly the same concentrations when operated at simulated altitudes (reduced exhaust pressures). Performance curves for the device while operating at atmospheric pressure are shown in Figure 7.

Information about carbon molecular sieves is scarce due to the proprietary status of the current manufacturing processes.[5,6,7] Carbon molecular sieves have demonstrated stability at high temperature and in acidic media, and a low affinity for water. They are produced by pyrolysis of many thermosetting polymers, such as, polyvinylidene chloride (PVDC), polyfurfuryl alcohol, cellulose triacetate, and saran copolymer. In contrast to zeolite molecular sieves which have a uniform pore size, carbon molecular sieves have a narrow pore size distribution. Further, this size distribution may be adjusted by changing the conditions of the manufacturing process.

In this work a carbon molecular sieve referred to as Takeda 3A was determined the most effective at separating oxygen and argon. This material is manufactured by Takeda Chemical Industries, Ltd., 12-10, Nihonbashi 2-chome, Chuo-ku, Tokyo 103, Japan. The material was provided to our laboratory by the U.S. representative for Takeda Chemical; TIGG Corporation, Box 11661, Pittsburgh, Pennsylvania. Although during this work the Takeda 3A carbon molecular sieve could only be obtained in limited quantities, this material is currently available in bulk quantities. The material was received as ~1/8 inch pellets (a typical pellet had a diameter of 2.36 mm and a length of 5.18 mm). These pellets were reduced in size in our laboratory by a Model No. 3383-L10 Wiley mill with a 10 mesh delivery unit. A mechanical sieving procedure was used to separate the mesh size desired. In general, mesh sizes of 10X40 and 16X40 were used in this work. Residual dust was removed by blowing compressed air through sieving screens containing the material. Size reduction was the only pretreatment performed before loading the material into the experimental apparatus. Size reduction was conducted to improve the mass transfer characteristics of the adsorbent. Experiments with different mesh sizes clearly indicated that the smaller mesh sizes are more effective at separating oxygen and argon.

Although the secondary oxygen purifier was capable of generating 99% purity oxygen, one disadvantage is the requirement for a feed gas with a concentration of approximately 95% oxygen and 5% argon. The next goal of this effort was to construct a new oxygen concentrator capable of generating 99% purity oxygen directly from compressed air. The approach was to devise a method for integrating the secondary oxygen purifier into a standard oxygen concentrator. The new device would have the capability of separating nitrogen and argon from compressed air.

99% PURITY MOLECULAR SIEVE OXYGEN GENERATOR

In 1989 a small-scale adsorption apparatus consisting of four interconnected adsorption beds, several valves, and an electronic timer demonstrated that oxygen concentrations of 99% could be achieved directly from compressed air using a PSA technique (Figure 8).[8,9] In Figure 8 adsorption beds A and B contained 585g of 16X40 mesh 5AMG zeolite molecular sieve and beds C and D contained 394g of 10X40 mesh carbon molecular sieve. The carbon molecular sieve was reduced in particle size from pellets having a diameter of ~2.36 mm to 10X40 mesh by the mechanical grinding procedure described previously. Adsorbent containment was achieved by four assemblies, each consisting of a metal screen, a foam pad, and a coil spring. Two assemblies were located at the inlets to beds A and B and two were at the outlets of beds C and D. The zeolite and carbon molecular sieve beds were connected in series. Hence, the gas flow passed sequentially from the zeolite molecular sieve bed to the carbon molecular sieve bed. The adsorption beds were constructed from 5.08 cm (2.00 in) OD stainless steel tubing. Beds A and B had a length of 43.1 cm and beds C and D had a length of 36.8 cm. The zeolite molecular sieve beds (A and B) were connected near their outlets by a 0.71 mm ID purge orifice. Valves V2-V7 were air operated valves manufactured by the Whitey Company (Part No. SS-92MA-NC). An electronic valve timer permitted adjustment of the cycle time of the apparatus. The device was fed dry compressed air with a concentration of 20.97% oxygen, 0.96% argon, and 78.07% nitrogen, as measured by a Perkin-Elmer medical gas analyzer (Model No. MGA-1100). The accuracy of the gas analyzer was ±0.1%.

During operation the apparatus was alternately cycled through steps of pressurization and depressurization in a manner similar to a standard oxygen concentrator. In the first half-cycle valves V2, V5, and V7 are activated open, while valves V3, V4, and V6 remain closed. Inlet air pressurizes beds A and C, and establishes a product flow at the outlet port of bed C. As the air passes through the adsorbent beds, nitrogen is preferentially adsorbed in bed A and argon is preferentially adsorbed in bed C. Hence, oxygen is concentrated and withdrawn as a product gas through valves V2 and V1. Simultaneously, bed B is regenerated by depressurization to the surrounding ambient pressure, countercurrent purging by a portion of the product flow from bed A, and countercurrent purging by a flow resulting from the partial depressurization of bed D. This depressurization exhausts the previously adsorbed nitrogen and argon to the ambient surroundings. During the second half-cycle valves V3, V4, and V6 are opened, while valves V2, V5, and V7 are closed. During this phase of the cycle beds B and D are pressurized and product gas is withdrawn through V3 and V1. Hence, in the second half-cycle beds B and D simply exchange roles with beds A and C. By repeating steps of adsorption and desorption, a continuous stream of very high purity oxygen is produced. Additionally, it should be noted that the apparatus does not require a purge flow for regeneration of the secondary or carbon molecular sieve adsorbent beds during the depressurization cycle. This feature improves the efficiency of the apparatus by reducing the feed air consumption.

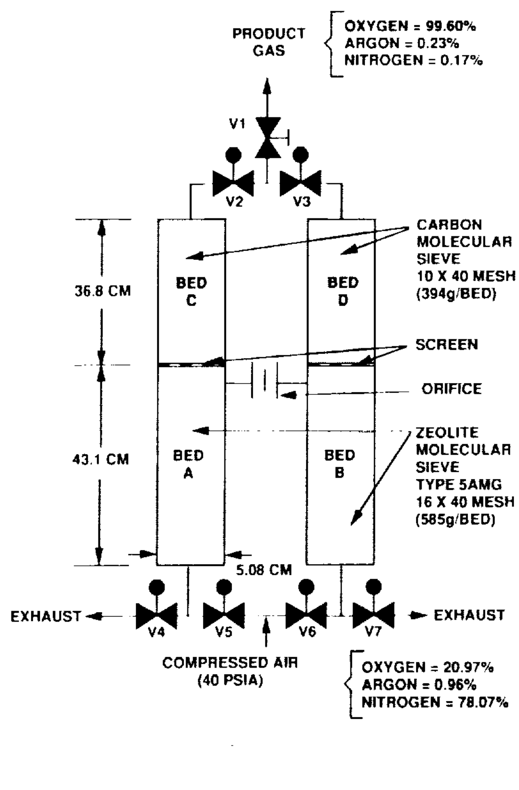


Figure 8. A Small-scale 99% Purity Molecular Sieve Oxygen Generator.

The effects of inlet air pressure on the oxygen purity produced by the small-scale 99% purity oxygen generator are given in Figure 9 and Table 1. The corresponding argon and nitrogen concentrations are shown in Figures 10 and 11. Clearly, the device is capable of producing 99% purity oxygen directly from compressed air. The maximum observed oxygen purity of 99.6% occurred at an inlet pressure of 275.8 KPa (40 psia). (The maximum oxygen purity has recently been increased to 99.7%.) At 241.3 KPa (35 psia) the oxygen purities at low product flows were high but decreased significantly as the product flow increased. At 344.8 KPa (50 psia) the device did not quite achieve 99% purity even at low product flows. Argon concentrations in the product gas decreased as the inlet pressure increased (Figure 10). Inlet air flows at pressures of 241.3 (35), 275.8 (40), 310.3 (45), and 344.8 KPa (50 psia) were 52, 61, 71, and 80 SLPM, respectively. Inlet air flow remained nearly constant with changes in product flow. This characteristic is also observed for standard oxygen generators.

Table 1. Oxygen Concentrations for the Small-Scale 99% Purity Oxygen Generator at a Cycle Time of 15 Seconds and Bed Temperature of 297K.

	Oxygen Concentration (%)			
	Inlet Pressure, KPa (psia)			
Product Flow (SLPM)	241.3 (35)	275.8 (40)	310.3 (45)	344.8 (50)
0.1	99.50	99.60	99.44	98.77
0.2	99.50	99.50	99.48	98.85
0.3	99.37	99.44	99.27	98.58
0.4	99.00	99.17	99.15	98.37
0.5	98.79	98.94	98.85	98.12
0.6	97.92	98.58	98.50	97.67
0.7	97.08	98.10	98.33	97.25
0.8	96.08	97.38	97.58	96.63
0.9	94.35	96.65	96.88	95.63
1.0	91.10	95.21	95.65	95.00

Oxygen recoveries were calculated based on Eqn. 1 and are given in Table 2 for purities of 99% and 95%. Recovery is a measure of how effectively oxygen is separated from the inlet air flow. Generally, oxygen recoveries are lower for small-scale oxygen concentrators when compared with full scale concentrators. The recoveries at 95% purity are similar to those obtained for other small-scale standard oxygen concentrators.[10]

$$\text{Oxygen Recovery (\%)} = \frac{m_p\, y_o}{m_i\,(0.21)}(100) \qquad (1)$$

Table 2. Oxygen Recovery at 99% and 95% Purity for the Small-scale 99% Purity Oxygen Generator.

	Oxygen Recovery (%)			
	Inlet Pressure, KPa (psia)			
	241.3 (35)	275.8 (40)	310.3 (45)	344.8 (50)
99% Purity	4.17	3.81	3.15	--
95% Purity	7.73	7.45	6.67	5.59

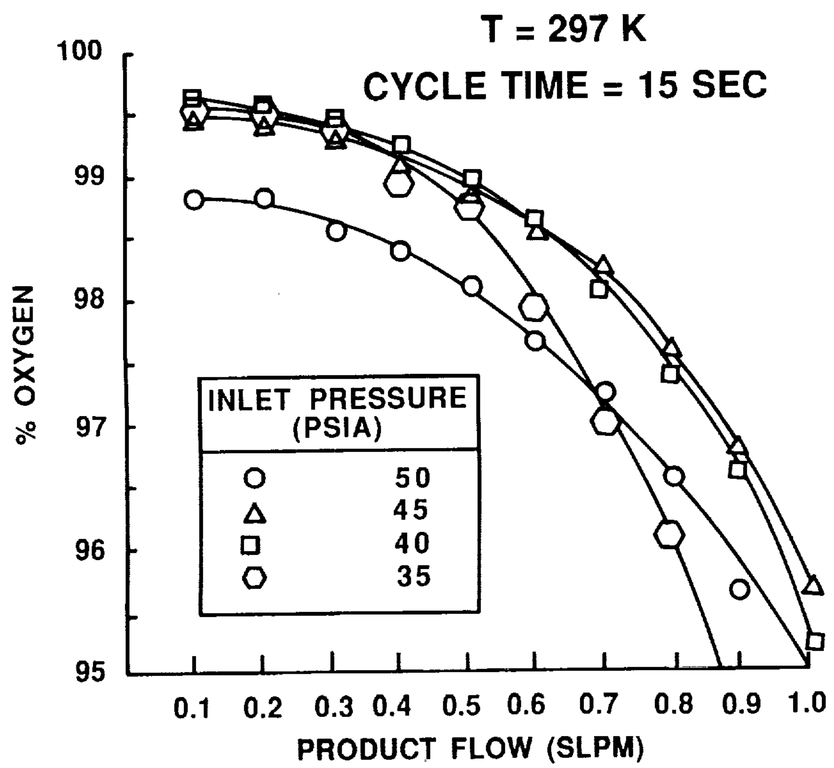


Figure 9. Oxygen Concentrations for the Small-Scale 99% Purity Oxygen Generator at Several Inlet Air Pressures.

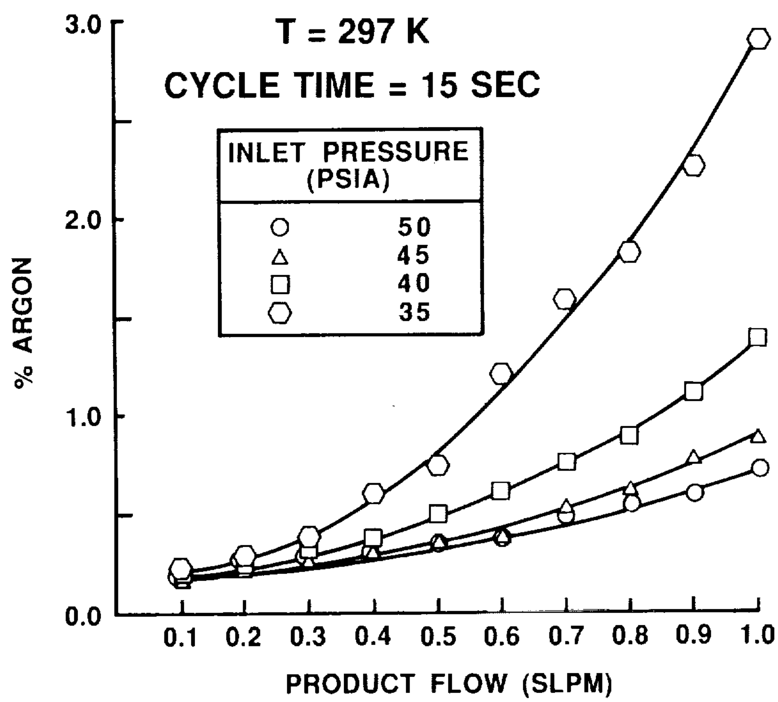


Figure 10. Argon Concentrations for the Small-Scale 99% Purity Oxygen Generator at Several Inlet Air Pressures.

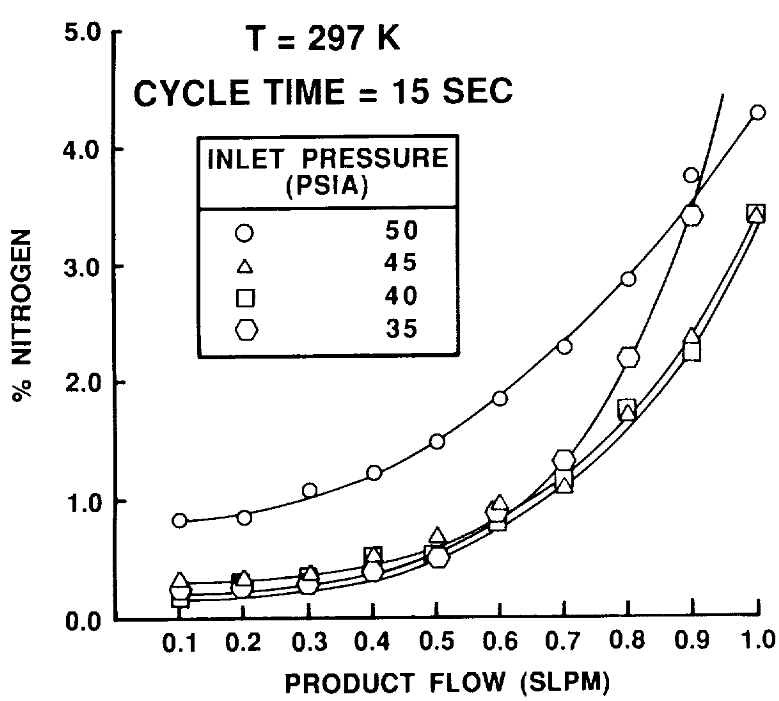


Figure 11. Nitrogen Concentrations for the Small-Scale 99% Purity Oxygen Generator at Several Inlet Air Pressures.

Operation of the oxygen concentrator at higher adsorbent bed temperatures causes a decrease in performance which is similar to that observed for standard oxygen generators.[11] This performance reduction occurs because molecular sieve adsorption capacity is reduced at higher temperatures. However, decreasing the cycle time of the concentrator will generally compensate for this reduced capacity.

Optimization studies of the 99% purity oxygen generator indicate a carbon to zeolite molecular sieve bed length ratio of 3/4 appears to maximize performance, assuming the diameter of the beds is equal.[12] Comparison of performance curves for 5AMG and OXYSIV-5 zeolite molecular sieves show increased performance with OXYSIV-5. The current highest oxygen productivity observed for the 99% molecular sieve oxygen generator is 0.33 (SLPM of 99% oxygen)/(Kg of total adsorbent).

TECHNOLOGY TRANSFER

The invention may be useful in any application where 99% purity oxygen is required or desired. Possible applications include filling of oxygen storage vessels or gas bottles, welding, glassblowing, industrial air separation processes, medical breathing systems, aircraft breathing systems, and space breathing systems. The invention would appear to be ideally suited to remote locations requiring 99% oxygen.

Further development of the technology is needed before the invention can be applied commercially. Presently, only small-scale laboratory devices have been constructed using this process. The next logical step in the development of the technology would be to construct a larger device with the capability of generating greater quantities of 99% oxygen. Armstrong Laboratory is considering two approaches toward commercialization. The first approach would be outright licensing of the invention. The second approach would involve a Cooperative Research and Development Agreement (CRDA) and licensing with a commercial firm. Any interested parties should contact Mr Douglas Blair, Armstrong Laboratory, Office of Research and Technology Applications (AL/XPPO), Brooks AFB, Texas (512-536-2838).

CONCLUSIONS

A new molecular sieve oxygen generator capable of generating oxygen purities of up to 99.7% directly from compressed air has been invented. The apparatus appears to have characteristics similar to standard oxygen generators but produces higher oxygen purities. The device may find use in aircraft and medical breathing systems, and industrial air separation systems.

NOMENCLATURE

m	= mass flow rate
psia	= pounds/square inch absolute
psig	= pounds/square inch gauge
SLPM	= standard liters/minute (referenced to 273K and 1 atm)
y	= mole fraction in the gas phase

Subscripts

i	= inlet
o	= oxygen
p	= product

ACKNOWLEDGMENT

The author wishes to thank Takeda Chemical Industries, Ltd., Tokyo, Japan and TIGG Corporation, Pittsburgh, Pennsylvania for supplying the carbon molecular sieve used in this work.

LITERATURE CITED

1. Tedor, J.B., and J.P. Clink, "Man Rating the B-1B Molecular Sieve Oxygen Generation System," Technical Report No. USAFSAM-TR-87-4, USAF School of Aerospace Medicine, Brooks AFB, Texas (1987).

2. Breck, D.W., Zeolite Molecular Sieves, John Wiley and Sons, Inc., New York, New York (1974).

3. Miller, G.W., and C.F. Theis, "Secondary Oxygen Purifier for Molecular Sieve Oxygen Concentrator," U.S. Patent No. 4,813,979 (1989).

4. Miller, G.W., and C.F. Theis, "Secondary Oxygen Purifier for Molecular Sieve Oxygen Concentrator," SAFE Journal, 19, 3, 27 (1989).

5. Walker, P.L., L.G. Austin, and S.P. Nandi, Chemistry and Physics of Carbon, ed. P.L. Walker, Marcel Dekker, N.Y., 2, 257 (1966).

6. Ma, Y.H., W. Sun, M. Bhandarkar, and G.W. Miller, "Adsorption and Diffusion of Oxygen, Nitrogen, Methane, and Argon in Molecular Sieve Carbons at Elevated Pressures," Separations Technology, 1, 90 (1991).

7. Ma, Y.H., W. Sun, M. Bhandarkar, and J. Wang, "Adsorption and Diffusion of Oxygen, Nitrogen, Methane, and Argon in Molecular Sieve Carbons," Technical Report No. USAFSAM-TR-89-32, USAF School of Aerospace Medicine, Brooks AFB, Texas (1990).

8. Miller, G.W., and C.F. Theis, "Molecular Sieve Oxygen Concentrator with Secondary Oxygen Purifier," U.S. Patent No. 4,880,443 (1989).

9. Miller, G.W., and C.F. Theis, "99% Purity Molecular Sieve Oxygen Concentrator," SAFE Journal, 20, 1, 6 (1990).

10. Theis, C.F., K.G. Ikels, and R.G. Dornes, "A Small Oxygen Concentrator," Technical Report No. USAFSAM-TR-85-18, USAF School of Aerospace Medicine, Brooks AFB, Texas (1985).

11. Miller, G.W., and C.F. Theis, "Thermal Testing of a 99% Purity Molecular Sieve Oxygen Concentrator," SAFE Journal, 21, 2, 26 (1991).

12. Miller, G.W., and C.F. Theis, "Optimization Studies on a 99% Purity Molecular Sieve Oxygen Concentrator: Effects of the Carbon to Zeolite Molecular Sieve Ratio," to be presented at the 1991 SAFE International Symposium, 11-13 November 1991, Riviera Hotel, Las Vegas, Nevada.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 1991	Conference Publication

4. TITLE AND SUBTITLE
Technology 2001
Volume 1

5. FUNDING NUMBERS

6. AUTHOR(S)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NASA Technology Utilization Division

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, DC 20546

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
NASA CP-3136

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified - Unlimited

Subject Category 99

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Proceedings from the Technical Sessions of the Technology 2001 Conference and Exposition, December 3-5, 1991, San Jose, CA. Volume 1 features 60 papers presented during 30 concurrent sessions.

14. SUBJECT TERMS
technology transfer
computer technology
materials sciencd
robotics

15. NUMBER OF PAGES
535

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev 2-89)
Prescribed by ANSI Std Z39-18
298-102

C-7